IBM SPSS Missing Values 19

Note: Before using this information and the product it supports, read the general information under Notices sur p. 98.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Valeurs manquantes fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Valeurs manquantes doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, est un des leaders dans le domaine des solutions logicielles d'analyse prédictive. Le portfolio complet des produits de la société — Data collection, Statistics, Modeling et Deployment — capture les opinions et les attitudes du public, prédit les résultats des interactions futures des clients, et agit ensuite sur ces données en intégrant les analyses dans les processus commerciaux. Les solutions SPSS Inc. répondent aux objectifs commerciaux interdépendants d'une organisation dans sa totalité en se concentrant sur la convergence des analyses, de l'architecture informatique et des processus commerciaux. Des clients issus du milieu des affaires, du milieu gouvernemental ou du milieu académique, dans le monde entier, font confiance à la technologie SPSS Inc., et la considère comme un atout pour attirer et retenir leurs clients, ou encore augmenter leur nombre, tout en réduisant les fraudes et les risques. SPSS Inc. a été acheté par IBM en octobre 2009. Pour plus d'informations, visitez le site *http://www.spss.com*.

Support technique

Un support technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits SPSS Inc. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, consultez le site Web SPSS Inc. à l'adresse *http://support.spss.com*, ou recherchez votre représentant local à la page *http://support.spss.com/default.asp?refpage=contactus.asp* Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Service clients

Si vous avez des questions concernant votre envoi ou votre compte, contactez votre bureau local, dont les coordonnées figurent sur le site Web à l'adresse : *http://www.spss.com/worldwide*. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

Séminaires de formation

SPSS Inc. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, contactez votre bureau local dont les coordonnées sont indiquées sur le site Web à l'adresse : *http://www.spss.com/worldwide*.

Documents supplémentaires

Les ouvrages *SPSS Statistics : Guide to Data Analysis*, *SPSS Statistics : Statistical Procedures Companion*, et *SPSS Statistics : Advanced Statistical Procedures Companion*, écrits par Marija Norušis et publiés par Prentice Hall, sont suggérés comme documentation supplémentaire. Ces publications présentent les procédures statistiques des modules SPSS Statistics Base, Advanced Statistics et Regression. Que vous soyez novice dans les analyses de données ou prêt à utiliser des applications plus avancées, ces ouvrages vous aideront à exploiter au mieux les fonctionnalités offertes par IBM® SPSS® Statistics. Pour obtenir des informations supplémentaires y compris le contenu des publications et des extraits de chapitres, visitez le site web de l'auteur : *http://www.norusis.com*

Contenu

Partie I:
Guide de l'utilisateur

Chapitre
1

Introduction aux valeurs manquantes

Les observations ayant des valeurs manquantes représentent un défi important car les procédures de modélisation classiques éliminent tout simplement ces observations des analyses. Lorsque les valeurs manquantes sont peu nombreuses (très approximativement, moins de 5% du nombre total d'observations) et que ces valeurs peuvent être considérées comme aléatoirement manquantes, c'est-à-dire qu'une valeur manquante ne dépend pas des autres valeurs, alors la méthode traditionnelle d'élimination est relativement "sûre". L'option Valeurs manquantes peut vous aider à déterminer si l'élimination est suffisante et vous proposer des méthodes de traitement des valeurs manquantes lorsqu'elle ne suffit pas.

Analyse des valeurs manquantes ou procédures à imputation multiple

L'option Valeurs manquantes propose deux ensembles de procédures permettant de traiter les valeurs manquantes :

- Les procédures d'Imputation multiple proposent des analyses de schémas de données manquantes, orientées vers une imputation multiple finale des valeurs manquantes. C'est-à-dire que plusieurs versions de l'ensemble de données sont produites, chacune d'elle contenant son propre ensemble de données imputées. Lorsque des analyses statistiques sont effectuées, les estimations de paramètre pour tous les ensembles de données imputés sont combinées ce qui génère des estimations généralement plus précises que celles provenant uniquement de l'imputation.
- L'Analyse des valeurs manquantes contient un ensemble légèrement différent d'outils descriptifs pour l'analyse de données manquantes (plus particulièrement le test MCAR Little) et comprend un grand nombre de méthodes d'imputation simple. Veuillez noter que l'imputation multiple est généralement considérée comme supérieure à l'imputation simple.

Tâches des valeurs manquantes

Vous pouvez commencer à analyser des valeurs manquantes en suivant ces étapes de base :

- **Examiner le caractère manquant.** Utilisez l'analyse des valeurs manquantes et Analyser les schémas pour explorer des schémas de valeurs manquantes dans vos données et déterminer si l'imputation multiple est nécessaire.
- **Imputer les valeurs manquantes.** Utilisez Imputer des valeurs de données manquantes pour imputer les valeurs manquantes.
- **Analyser les données "complètes".** Utilisez n'importe quelle procédure prenant en charge les données à imputation multiple. Consultez Analyse de données à imputation multiple sur p. 30

pour obtenir des informations sur l’analyse des ensembles de données à imputation multiple et sur une liste de procédures prenant en charge ces données.

Chapitre
2

Analyse des valeurs manquantes

La procédure d'analyse de la valeur manquante exécute trois fonctions principales :

- Elle décrit le type des données manquantes. Quel est l'emplacement des valeurs manquantes ? Quelle est l'importance de leur nombre ? Les paires de variables ont-elles tendance à contenir des valeurs manquantes dans les observations multiples ? Les données ont-elles des valeurs extrêmes ? Les valeurs manquent-elles de façon aléatoire ?
- Estime les moyennes, écarts-types, covariances et corrélations pour différentes méthodes relatives aux valeur manquantes : par liste, par pair, régression ou EM (prévision-maximisation). La méthode concernant seulement les composantes non valides affiche également l'effectif des observations complètes par paires.
- Remplit (impute) les valeurs manquantes avec des valeurs estimées à l'aide de méthodes de régression ou EM ; mais les résultats de l'imputation multiple sont généralement considérés comme plus précis.

L'analyse des valeurs manquantes vous aide à aborder de nombreux problèmes causés par des données incomplètes. Si des observations avec valeurs manquantes sont systématiquement différentes d'observations sans valeurs manquantes, cela peut aboutir à des résultats erronés. De même, les données manquantes peuvent réduire la précision des statistiques calculées car l'information disponible est inférieure à celle initialement prévue. Un autre problème est que les hypothèses effectuées en aval de nombreuses procédures statistiques sont basées sur des observations complètes et que les valeurs manquantes peuvent compliquer la théorie requise.

Exemple : Lors de l'évaluation d'un traitement contre la leucémie, plusieurs variables sont mesurées. Cependant, toutes les mesures différentes ne sont pas disponibles pour chaque patient. Le type des données manquantes est affiché, mis en tableau et s'avère être aléatoire. Une analyse EM est utilisée afin d'estimer les moyennes, les corrélations et les covariances. Elle permet également de déterminer si les données sont des valeurs manquantes complètement aléatoires. Les valeurs manquantes sont remplacées par des valeurs imputées et enregistrées dans un nouveau fichier de données pour des analyses supplémentaires.

Statistiques : Statistiques univariées incluant le nombre de valeurs non manquantes, la moyenne et l'écart-type, et le nombre de valeurs manquantes et de valeurs extrêmes. Moyennes estimées, matrice de covariance, matrice de corrélation déterminées à l'aide des méthodes de type toutes observations incomplètes, seulement les composantes non valides, des méthodes EM ou de régression. Le test MCAR avec les résultats EM. Récapitulatif des moyennes par différentes méthodes. Pour les groupes définis par des valeurs manquantes par opposition à ceux définis par des valeurs non manquantes : Tests *T*. Pour toutes les variables : modèles des valeurs manquantes affichées observations-par-variable.

Analyse des données

Données. Les données peuvent être nominales ou quantitatives (échelle ou continues). Toutefois, vous ne pouvez estimer les statistiques et imputer les données manquantes que pour les variables quantitatives. Pour chaque variable, les valeurs manquantes qui ne sont pas codées comme Manquantes système doivent être définies comme Manquantes utilisateur. Par exemple, si dans un questionnaire, l'un des éléments a pour réponse *Ne sais pas*, que cette réponse est codée par le chiffre 5 et que vous souhaitez traiter cette réponse comme manquante, l'élément concerné se verra alors attribuer 5 comme valeur manquante utilisateur.

Pondérations d'effectif. Cette procédure utilise les pondérations d'effectifs (réplication). Les observations ayant une valeur de pondération de réplication négative ou nulle sont ignorées. Les pondérations non entières sont tronquées.

Hypothèses : L'estimation selon l'exclusion de toute observation incomplète, l'exclusion seulement des paires non valides ou l'exclusion de régression sont basées sur l'hypothèse que le motif des valeurs manquantes ne dépend pas des valeurs des données. (Cette condition est connue sous le terme **Valeur manquante complètement aléatoire** ou MCAR.) Par conséquent, toutes les méthodes d'estimation (y compris la méthode EM) donnent des estimations cohérentes et non biaisées des corrélations et des covariances lorsque les données sont de type MCAR. La violation de l'hypothèse MCAR peut aboutir à des estimations biaisées produites par les méthodes de régression, de type toutes observations incomplètes ou de type seulement les composantes non valides. Si les données ne sont pas de type MCAR, vous devez utiliser l'estimation EM.

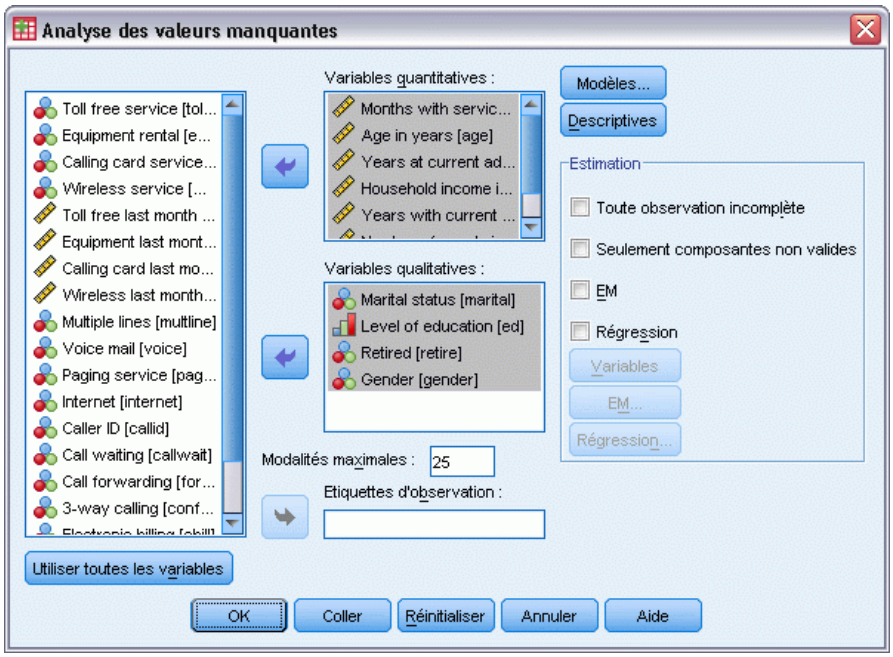
Les estimations EM sont basées sur l'hypothèse que le modèle des données manquantes est uniquement lié aux données observées. (Cette condition est appelée **valeur manquante aléatoire**, ou MAR.) Cette hypothèse permet d'ajuster les estimations à l'aide des informations disponibles. Par exemple, dans une enquête portant sur les études et le revenu, il est possible que les sujets ayant un bas niveau d'études présentent davantage de valeurs de revenu manquantes. Dans ce cas, les données sont de type MAR, au lieu de MCAR. En d'autres termes, pour le type MAR, la probabilité que le revenu soit enregistré dépend du niveau d'études du sujet. La probabilité peut varier en fonction du niveau d'études, mais pas en fonction du revenu *au sein de chaque niveau d'études*. Si la probabilité d'enregistrement du revenu varie aussi en fonction de la valeur du revenu dans chaque niveau d'études (par exemple, les personnes qui ont des revenus élevés sont susceptibles de ne pas les indiquer), les données ne sont ni de type MCAR, ni de type MAR. Cette situation n'est pas rare et, lorsqu'elle se présente, aucune des méthodes n'est appropriée.

Procédures apparentées : De nombreuses procédures vous permettent d'utiliser l'estimation de type toutes observations incomplètes ou de type seulement les composantes non valides. L'analyse de régression et facteur linéaires autorise le remplacement des valeurs manquantes par les valeurs moyennes. Dans le module complémentaire Prévisions, plusieurs méthodes sont disponibles afin de remplacer les valeurs manquantes en séries chronologiques.

Pour obtenir une analyse des valeurs manquantes

▶ A partir des menus, sélectionnez :
Analyse > Analyse des valeurs manquantes

Figure 2-1
Boîte de dialogue Analyse des valeurs manquantes



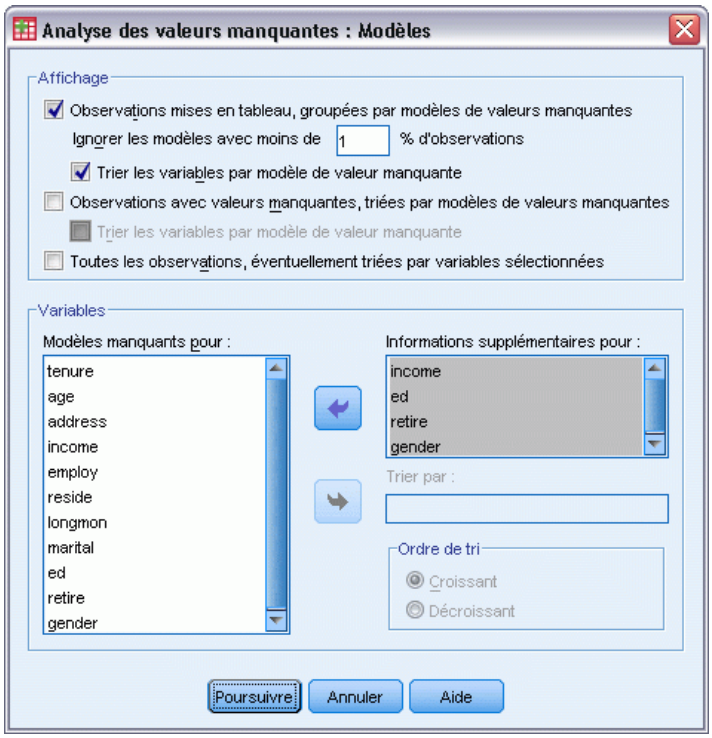
► Sélectionnez au moins une variable quantitative (échelle) pour l'estimation des statistiques et, éventuellement, pour l'imputation des valeurs manquantes.

Sinon, vous pouvez :

- Sélectionner des variables qualitatives (numériques ou chaîne) et entrer une limite relative au nombre de modalités (Modalités maximales).
- Cliquez sur Modèles pour mettre en tableau les patrons de données manquantes. Pour plus d'informations, reportez-vous à la section Affichage des patrons de valeurs manquantes sur p. 6.
- Cliquez sur Descriptives pour afficher les statistiques descriptives des valeurs manquantes. Pour plus d'informations, reportez-vous à la section Affichage des statistiques descriptives des valeurs manquantes sur p. 8.
- Sélectionnez une méthode d'estimation des statistiques (moyennes, covariances et corrélations) et, éventuellement, d'imputation des valeurs manquantes. Pour plus d'informations, reportez-vous à la section Estimation des statistiques et imputation des valeurs manquantes sur p. 9.
- Si vous sélectionnez EM ou Régression, cliquez sur Variables... pour spécifier le sous-ensemble à utiliser pour l'estimation. Pour plus d'informations, reportez-vous à la section Variables dépendantes et variables prédites sur p. 13.
- Sélectionnez une variable d'étiquette d'observation. Cette variable permet d'étiqueter les observations dans les tableaux de patrons qui affichent des observations individuelles.

Affichage des patrons de valeurs manquantes

Figure 2-2
Boîte de dialogue Modèles d'analyses des valeurs manquantes



Vous pouvez choisir d'afficher différents tableaux montrant les patrons et l'étendue des données manquantes. Ces tableaux vous permettent d'identifier :

- L'emplacement des valeurs manquantes.
- Si les paires de variables ont tendance à contenir des valeurs manquantes dans les observations individuelles.
- Si les valeurs de données sont extrêmes.

Affichage

Trois types de tableaux sont disponibles pour l'affichage des patrons de données manquantes.

Observations mises en tableau. Les patrons de valeurs manquantes dans les variables d'analyse sont mis en tableau, avec affichage des fréquences pour chaque patron. Utilisez l'option Trier les variables par modèle de valeur manquante pour indiquer si les effectifs et les variables sont triés selon la similarité des patrons. Utilisez l'option Omettez les modèles avec moins de n % d'observation pour éliminer les patrons qui se produisent rarement.

Observations avec valeurs manquantes. Chaque observation contenant une valeur manquante ou extrême est mise en tableau pour chaque variable d'analyse. Utilisez l'option Trier les variables par modèle de valeur manquante pour indiquer si les effectifs et les variables sont triés selon la similarité des patrons.

Toutes les observations. Chaque observation est mise en tableau, avec indication des valeurs manquantes et extrêmes pour chaque variable. Les observations sont listées suivant l'ordre dans lequel elles apparaissent dans le fichier de données, à moins qu'une variable de tri ne soit spécifiée dans Trier par.

Les symboles suivants sont utilisés dans les tableaux qui affichent des observations individuelles :

+	Valeur extrêmement haute
-	Valeur extrêmement basse
S	Valeur manquante par défaut
A	Premier type de valeur manquante utilisateur
B	Second type de valeur manquante utilisateur
C	Troisième type de valeur manquante utilisateur

Variables

Vous pouvez afficher des informations supplémentaires sur les variables incluses dans l'analyse. Les variables que vous ajoutez à l'option Informations supplémentaires pour apparaissent séparément dans le tableau des patrons manquants. Pour les variables quantitatives (échelle), c'est la moyenne qui apparaît ; dans le cas des variables qualitatives, il s'agit du nombre d'observations correspondant à un type dans chacune des modalités.

- **Trier par.** Les observations sont listées selon l'ordre croissant ou décroissant des valeurs de la variable spécifiée. Uniquement disponible pour Toutes les observations.

Pour spécifier les types de valeurs manquantes

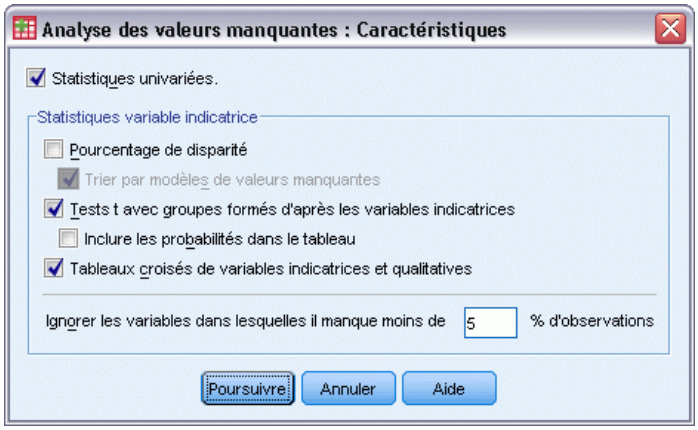
▶ Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez afficher les patrons de valeurs manquantes.

▶ Cliquez sur Modèles.

▶ Sélectionnez les tableaux de patron à afficher.

Affichage des statistiques descriptives des valeurs manquantes

Figure 2-3
Analyse des valeurs manquantes Boîte de dialogue Descriptives



Statistiques univariées

Les statistiques univariées vous permettent d'identifier l'étendue générale des données manquantes. Pour chaque variable, les éléments suivants apparaissent :

- Nombre de valeurs non manquantes
- Nombre et pourcentage de valeurs manquantes

Pour les variables quantitatives (échelle), les éléments suivants apparaissent également :

- Moyenne
- Ecart type
- Nombre de valeurs extrêmement élevées et basses

Statistiques variable indicatrice

Pour chaque variable, une variable indicatrice est créée. Cette variable qualitative indique si la variable est présente ou manquante pour une observation individuelle. Les variables indicatrices permettent de créer la disparité, le test *t* et les tableaux de fréquences.

Pourcentage de disparité. Affiche, pour chaque paire de variables, le pourcentage d'observations pour lesquelles une variable a une valeur manquante tandis que l'autre variable a une variable non manquante. Dans le tableau, chaque élément diagonal contient le pourcentage des valeurs manquantes pour une seule variable.

t tests avec groupes formés d'après les variables d'indication. Les moyennes de deux groupes sont comparées pour chaque variable quantitative, en utilisant les statistiques *t* de Student. Les groupes indiquent si une variable est présente ou manquante. Les statistiques *t*, les degrés de liberté, les effectifs des valeurs manquantes ou non manquantes et les moyennes des deux groupes sont affichés. Vous pouvez également afficher toutes les probabilités bilatérales associées aux statistiques *t*. Si l'analyse aboutit à au moins deux tests, n'utilisez pas ces probabilités pour tester la signification. Les probabilités ne sont appropriées que lorsqu'un seul test est calculé.

Mises en tableau croisés de variables d'indication et nominales. Un tableau est affiché pour chaque variable qualitative. Pour chacune des modalités, le tableau montre la fréquence et le pourcentage des valeurs non manquantes pour les autres variables. Les pourcentages de chaque type de valeur manquante sont également affichés.

Omettez les variables pour lesquelles il manque moins de n % d'observations. Pour réduire la dimension des tableaux, vous pouvez omettre les statistiques qui ne sont calculées que pour un petit nombre d'observations.

Pour afficher des statistiques descriptives

- Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez afficher les statistiques descriptives des valeurs manquantes.
- Cliquez sur Descriptives.
- Sélectionnez les statistiques descriptives à afficher.

Estimation des statistiques et imputation des valeurs manquantes

Vous pouvez estimer les moyennes, les écarts-types, les covariances et les corrélations à l'aide des méthodes de type toutes observations incomplètes, de type seulement les composantes non valides, EM (prévision-maximisation) et/ou de régression. Vous pouvez également imputer les valeurs manquantes (valeurs de remplacement d'estimation). Notez que l'Imputation multiple est généralement considérée comme supérieure à l'imputation simple pour résoudre le problème des valeurs manquantes. Le test MCAR Little reste utile pour déterminer si l'imputation est nécessaire.

Méthode de type toutes observations incomplètes

Cette méthode utilise uniquement des observations complètes. Si l'une des variables d'analyse comprend des valeurs manquantes, l'observation est exclue du calcul.

Méthode de type seulement les composantes non valides

Cette méthode considère les paires de variables d'analyse et n'utilise une observation que si elle possède des valeurs non manquantes pour les deux variables. Les fréquences, les moyennes et les écarts-types sont calculés séparément pour chaque paire. Etant donné que les autres valeurs manquantes dans l'observation sont ignorées, les corrélations et les covariances pour deux variables ne dépendent pas des valeurs faisant défaut dans les autres variables.

Méthode EM

Cette méthode suppose une distribution pour les données partiellement manquantes et base les inférences sur la probabilité sous cette distribution. Chaque itération se compose d'une étape E et d'une étape M. L'étape E recherche la prévision conditionnelle des données « manquantes », en fonction des valeurs observées et des estimations en cours des paramètres. Ces prévisions sont ensuite substituées aux données « manquantes ». Dans l'étape M, les estimations du maximum de vraisemblance des paramètres sont calculées comme si les données manquantes avaient été

remplies. Le terme « manquantes » est indiqué entre guillemets, car les valeurs manquantes ne sont pas directement remplies. En fait, certaines de leurs fonctions sont utilisées dans le log-vraisemblance.

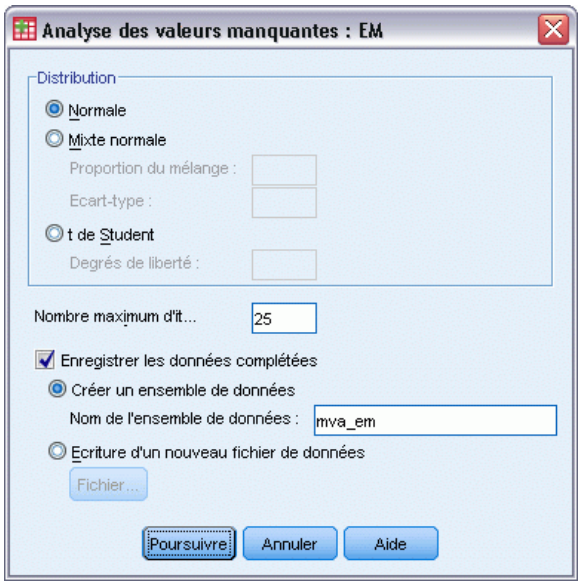
La statistique du Khi-deux de Roderick J. A. Little, qui permet de tester si les valeurs sont de type valeur manquante complètement aléatoire (MCAR), apparaît sous la forme d'une note de bas de page dans les matrices EM. Pour ce test, l'hypothèse nulle est que les données sont de type MCAR et la valeur *p* est significative au niveau 0,05. Si la valeur est inférieure à 0,05, les données ne sont pas des valeurs manquantes complètement aléatoires. Les données peuvent être de type MAR ou NMAR (valeur non manquante aléatoire). Vous ne pouvez pas supposer l'un ou l'autre type et devez analyser les données pour déterminer dans quelle mesure elles sont manquantes.

Méthode de régression :

Cette méthode calcule plusieurs estimations de régression linéaire et permet d'augmenter les estimations à l'aide de composants aléatoires. A chaque valeur prévue, la procédure peut ajouter un résidu à partir d'une observation complète sélectionnée aléatoirement, un écart normal aléatoire ou un écart aléatoire (redimensionné par la racine carrée du carré moyen résiduel) à partir de la distribution *t*.

Options de l'estimation EM

Figure 2-4
Analyse des valeurs manquantes. Boîte de dialogue EM.



En utilisant un processus itératif, la méthode EM estime la moyenne, la matrice de covariance et la corrélation des variables quantitatives (échelle) présentant des valeurs manquantes.

Distribution : La méthode EM effectue des inférences basées sur la vraisemblance sous la distribution spécifiée. Par défaut, une distribution normale est supposée. S'il est établi que les extrémités de la distribution sont plus allongées que celles d'une distribution normale, vous pouvez demander que la procédure construise la fonction de vraisemblance à partir d'une distribution *t* de

Student avec *n* degrés de liberté. En outre, la distribution mixte normale fournit une distribution avec des extrémités plus longues. Spécifiez le ratio des écarts-types de la distribution mixte normale et la proportion du mélange des deux distributions. La distribution mixte normale suppose que seuls les écarts-types des distributions diffèrent. Les moyennes doivent être les mêmes.

Nombre maximum d'itérations : Fixe le nombre maximum d'itérations pour estimer la véritable covariance. La procédure s'arrête lorsque ce nombre d'itérations est atteint, même si les estimations n'ont pas convergé.

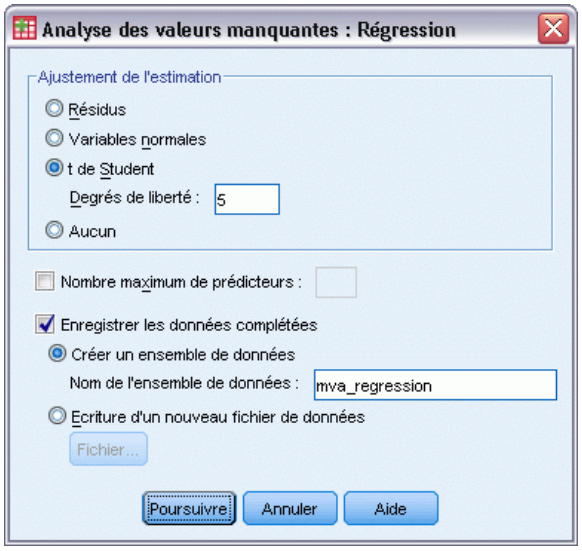
Enregistre les données complétées. Vous pouvez enregistrer un ensemble de données avec les valeurs imputées à la place des valeurs manquantes. Toutefois, gardez à l'esprit que les statistiques basées sur la covariance qui utilisent les valeurs imputées sous-estimeront leurs valeurs de paramètre respectives. Le degré de sous-estimation est proportionnel au nombre d'observations non observées conjointement.

Spécifier les options EM

- Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez estimer les valeurs manquantes à l'aide de la méthode EM.
- Sélectionnez EM dans le groupe Estimation.
- Pour spécifier les variables dépendantes (prévues) et explicatives, cliquez sur Variables. Pour plus d'informations, reportez-vous à la section Variables dépendantes et variables prédites sur p. 13.
- Cliquez sur EM.
- Sélectionnez les options EM souhaitées.

Options de l'estimation de la régression

Figure 2-5
Analyse des valeurs manquantes. Boîte de dialogue Régression



La méthode de régression estime les valeurs manquantes à l'aide de plusieurs régressions linéaires. La moyenne, la matrice de covariance et la matrice de corrélation des prévisions sont affichées.

Ajustement de l'estimation. La méthode de régression peut ajouter un composant aléatoire aux estimations de la régression. Vous pouvez sélectionner résidus, normales, *t* de Student ou aucun ajustement.

- **Résidus.** Les termes d'erreur sont choisis de manière aléatoire à partir des résidus observés de l'ensemble des observations à ajouter aux estimations de la régression.
- **Normale.** Les termes d'erreur sont choisis de manière aléatoire à partir d'une distribution de valeur théorique 0 et d'écart-type égal à la racine carrée du terme d'erreur sur la moyenne des carrés de la régression.
- **t de Student.** Les termes d'erreur sont choisis de manière aléatoire à partir de la distribution t(n), et redimensionnés par l'erreur sur la racine de la moyenne des carrés (RMSE).

Nombre maximum de prédicteurs. Fixe une limite maximale pour le nombre de variables prédites (indépendantes) utilisées dans le processus d'estimation.

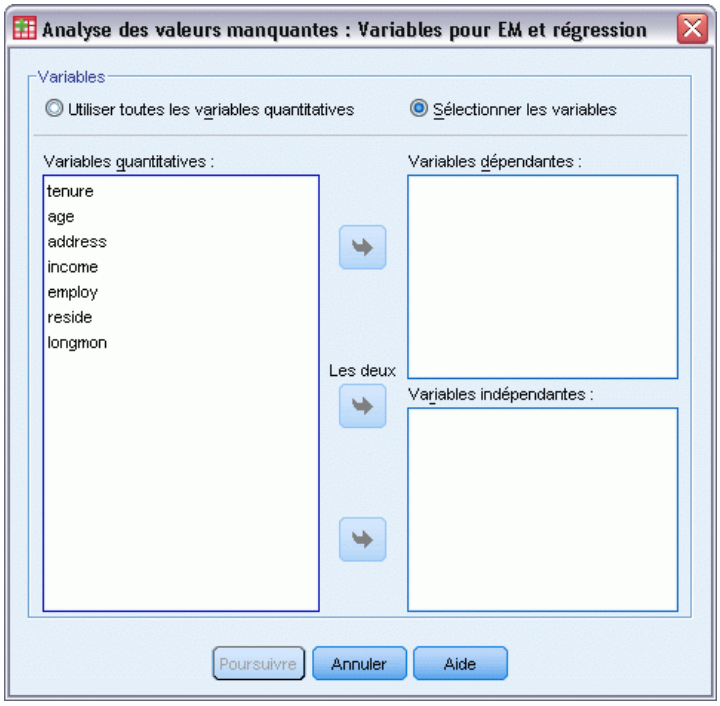
Enregistre les données complétées. Ecrit un ensemble de données dans la session en cours ou dans un fichier de données externe IBM® SPSS® Statistics, avec les valeurs manquantes remplacées par des valeurs estimées via la méthode de régression.

Spécifier les options de régression

- ▶ Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez estimer les valeurs manquantes à l'aide de la méthode de régression.
- ▶ Sélectionnez Régression dans le groupe Estimation.
- ▶ Pour spécifier les variables dépendantes (prévues) et explicatives, cliquez sur Variables. Pour plus d'informations, reportez-vous à la section Variables dépendantes et variables prédites sur p. 13.
- ▶ Cliquez sur Régression.
- ▶ Sélectionnez les options de régression souhaitées.

Variables dépendantes et variables prédites

Figure 2-6
Analyse des valeurs manquantes. Boîte de dialogue Variables pour EM et Régression



Par défaut, toutes les variables quantitatives sont utilisées pour l'estimation par EM et régression. Le cas échéant, vous pouvez choisir des variables spécifiques en tant que variables dépendantes et variables prédites dans les estimations. Une variable donnée peut figurer dans les deux listes ; cependant, dans certaines circonstances, vous pouvez être amené à limiter l'utilisation d'une variable. Par exemple, certains analystes trouvent inconfortable d'estimer les valeurs des variables de sortie. Il se peut également que vous préfériez utiliser différentes variables pour différentes estimations et exécuter la procédure plusieurs fois. Par exemple, si un ensemble d'éléments contient les évaluations des infirmières et un autre les évaluations des médecins, vous pouvez être amené à lancer un traitement à l'aide des éléments des infirmières pour estimer les éléments manquants des infirmières et un autre pour estimer les éléments des médecins.

L'utilisation de la méthode de régression soulève un autre point. Dans la régression multiple, l'utilisation d'un sous-ensemble volumineux de variables indépendantes peut générer des valeurs prévues moins pertinentes que celles produites par un sous-ensemble plus petit. Par conséquent, une variable ne peut être utilisée que si elle atteint une limite *F* pour introduire de 4,0. Cette limite peut être modifiée à l'aide d'une syntaxe.

Pour spécifier les variables dépendantes et les variables prédites

- Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez estimer les valeurs manquantes à l'aide de la méthode de régression.
- Sélectionnez EM ou Régression dans le groupe Estimation.

- Cliquez sur Variables.
- Si vous souhaitez utiliser des variables spécifiques, plutôt que la totalité des variables, en guise de variables dépendantes et de variables prédites, sélectionnez Sélectionner les variables, puis déplacez les variables vers les listes appropriées.

Commande MVA. Descriptives additionnelles

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier différentes variables descriptives pour les types de valeur manquante, les types de données et les types mis en tableau à l'aide du mot-clé `DESCRIBE` dans les sous-commandes `MPATTERN`, `DPATTERN` ou `TPATTERN`.
- Spécifier plusieurs variables de tri pour le tableau de types de données à l'aide de la sous-commande `DPATTERN`.
- Spécifier plusieurs variables de tri pour les types de données à l'aide de la sous-commande `DPATTERN`.
- Spécifier la tolérance et la convergence à l'aide de la sous-commande `EM`.
- Spécifier la tolérance et *F* pour introduire à l'aide de la sous-commande `REGRESSION`.
- Spécifier différentes listes de variables pour les paramètres EM et Régression via les sous-commandes `EM` et `REGRESSION`.
- Spécifier différents pourcentages en vue de supprimer les observations affichées pour chaque paramètre `TTESTS`, `TABULATE` et `MISMATCH`.

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Chapitre
3

Imputation multiple

Le but de l'imputation multiple est de générer des valeurs possibles pour les valeurs manquantes et de créer ainsi plusieurs ensembles de données "complets". Les procédures analytiques qui utilisent des ensembles de données à imputation multiple produisent des résultats pour chaque ensemble de données "complet" en plus de résultats combinés qui évaluent quels auraient été les résultats si l'ensemble de données d'origine ne contenait pas de valeurs manquantes. Ces résultats combinés sont généralement plus précis que ceux des méthodes d'imputation simple.

Variables d'analyse. Les variables d'analyse peuvent être :

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.
- **Echelle.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

  La procédure considère que le niveau de mesure approprié a été assigné à toutes les variables, bien que vous puissiez changer provisoirement le niveau de mesure d'une variable en cliquant avec le bouton droit de la souris sur la variable dans la liste des variables source, puis en sélectionnant un niveau de mesure dans le menu contextuel.

Dans la liste des variables, une icône indique le niveau de mesure et le type de données :

Niveau de mesure	**Le type de données**			
	Numérique	Chaîne	Date	Heure
Echelle (continue).		n/a		
Ordinales				
Nominales				

Pondérations d'effectif. Cette procédure utilise les pondérations d'effectifs (réplication). Les observations ayant une valeur de pondération de réplication négative ou nulle sont ignorées. Les pondérations non entières sont arrondies à l'entier le plus proche.

Pondération d'analyse. Les pondérations (de régression ou d'échantillon) d'analyse sont intégrées aux récapitulatifs des valeurs manquantes et aux modèles d'imputation appropriés. Les observations ayant une pondération d'analyse négative ou nulle sont exclues.

Echantillonnage. La procédure d'Imputation multiple ne traite pas de manière explicite les strates, les classes ou les autres structures d'échantillon complexes, bien qu'elle puisse accepter les pondérations d'échantillons finales sous la forme de variable de pondération d'analyse. Remarque : actuellement, les prodécures d'échantillonage complexe n'analysent pas de manière automatique les ensembles de données à imputation multiple. Pour une liste complète des procédures prenant en charge le regroupement, reportez-vous à Analyse de données à imputation multiple sur p. 30.

Valeurs manquantes : Les valeurs manquantes utilisateur et par défaut sont traitées comme des valeurs non valides, c'est-à-dire que ces deux types de valeurs manquantes sont remplacés lorsque des valeurs sont imputées et les deux sont traités comme valeurs non valides de variables utilisées comme variables prédites dans les modèles d'imputation. Les valeurs manquantes utilisateur et par défaut sont également traitées comme manquantes dans les analyses de valeurs manquantes.

Réplication de résultats (Imputer des valeurs de données manquantes). Si vous souhaitez répliquer exactement vos résultats d'imputation, outre les mêmes paramètres de procédure, utilisez la même valeur d'initialisation pour le générateur de nombres aléatoires, le même ordre de données et le même ordre de variables.

- **Génération de nombres aléatoires.** La procédure utilise la génération de nombres aléatoires pendant le calcul des valeurs imputées. Pour reproduire les mêmes résultats aléatoires à l'avenir, utilisez la même valeur d'initialisation pour le générateur de nombres aléatoires avant chaque exécution de la procédure d'imputation des valeurs de données manquantes.
- **Tri par observation.** Les valeurs sont imputées suivant l'ordre des observations.
- **Ordre des variables.** La méthode d'imputation à spécification entièrement conditionnelle (FCS) impute des valeurs dans l'ordre spécifié dans la liste Variables d'analyse.

Il existe deux boîtes de dialogue associées à l'imputation multiple.

- Analyser les modèles contient des mesures descriptives des modèles de valeurs manquantes dans les données et peut servir d'étape d'exploration avant l'imputation.
- Imputer les valeurs de données manquantes permet de générer des imputations multiples. Les ensembles de données complets peuvent être analysés avec des procédures prenant en charge des ensembles de données à imputation multiple. Consultez Analyse de données à imputation multiple sur p. 30 pour obtenir des informations sur l'analyse des ensembles de données à imputation multiple et sur une liste de procédures prenant en charge ces données.

Analyser les modèles

Analyser les modèles contient des mesures descriptives des modèles de valeurs manquantes dans les données et peut servir d'étape d'exploration avant l'imputation.

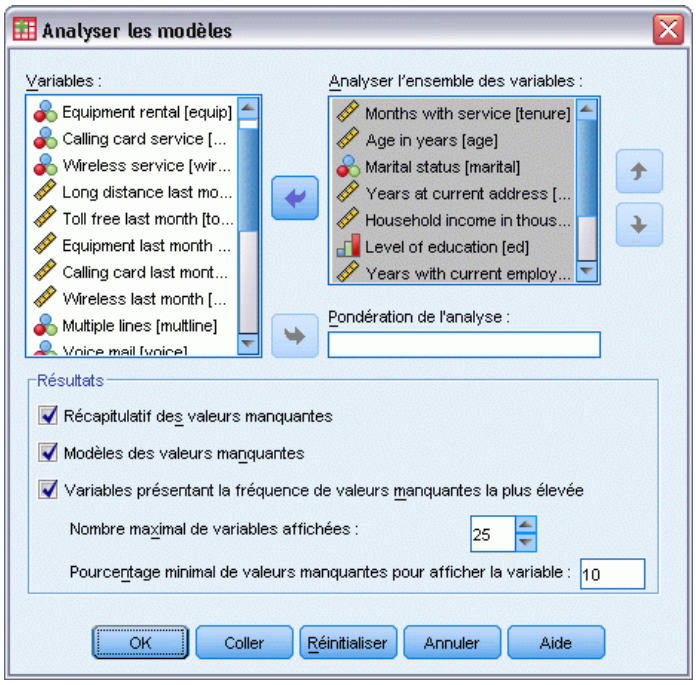
Exemple : Un fournisseur de services de télécommunication souhaite mieux comprendre les types d'utilisation des services dans sa base de données client. Il dispose de données complètes sur les services utilisés par les clients, mais les informations démographiques collectées par l'entreprise comportent certaines valeurs manquantes. L'analyse des modèles des valeurs manquantes peut contribuer à déterminer les étapes suivantes de l'imputation. Pour plus d'informations, reportez-vous à la section Utilisation de l'imputation multiple pour compléter et analyser un ensemble de données dans le chapitre 5 sur p. 52.

Pour analyser les modèles de données manquantes

A partir des menus, sélectionnez :

Analyse > Imputation multiple > Analyser les modèles...

Figure 3-1
Boîte de dialogue Analyser les modèles



- Sélectionnez au moins deux variables d'analyse. La procédure analyse les modèles de données manquantes pour ces variables.

Paramètres facultatifs

Pondération d'analyse. Cette variable contient des pondérations (de régression ou d'échantillon) d'analyse. La procédure intègre des pondérations d'analyse aux récapitulatifs des valeurs manquantes. Les observations ayant une pondération d'analyse négative ou nulle sont exclues.

Résultats. Le résultat facultatif suivant est disponible :

- **Récapitulatif des valeurs manquantes.** Il affiche un diagramme de panels en secteurs qui indique le nombre et le pourcentage de variables d'analyse, d'observations ou de valeurs de données individuelles qui contiennent une ou plusieurs valeurs manquantes.
- **Modèles de valeurs manquantes.** Permet d'afficher des modèles mis en tableau de valeurs manquantes. Chaque modèle correspond à un groupe d'observations avec le même modèle de données complètes et incomplètes dans les variables d'analyse. Vous pouvez utiliser ces résultats pour déterminer si la méthode d'imputation monotone peut être utilisée pour vos données, et dans le cas contraire, si vos données sont proches d'un modèle monotone. La procédure ordonne les variables d'analyse pour révéler ou ressembler à un modèle monotone. Si aucun modèle monotone n'existe après la réorganisation, vous pouvez en conclure que les données ont un modèle monotone lorsque les variables d'analyse sont ordonnées ainsi.
- **Variables avec l'effectif le plus élevé de valeurs manquantes.** Affiche un tableau des variables d'analyse triées par pourcentage de valeurs manquantes dans l'ordre décroissant. Ce tableau comprend des statistiques descriptives (moyenne et écart-type) pour les variables d'échelle.

  Vous pouvez contrôler le nombre de variables maximum à afficher et le pourcentage minimum manquant pour une variable à afficher. L'ensemble des variables qui répondent aux deux critères est affiché. Par exemple, définir le nombre de variables maximum sur 50 et le pourcentage minimum manquant sur 25 demande que le tableau affiche jusqu'à 50 variables ayant au moins 25% de valeurs manquantes. S'il existe 60 variables d'analyse mais que 15 seulement ont 25% ou plus de valeurs manquantes, le résultat ne comprendra que 15 variables.

Imputer les valeurs de données manquantes

Imputer les valeurs de données manquantes permet de générer des imputations multiples. Les ensembles de données complets peuvent être analysés avec des procédures prenant en charge des ensembles de données à imputation multiple. Consultez Analyse de données à imputation multiple sur p. 30 pour obtenir des informations sur l'analyse des ensembles de données à imputation multiple et sur une liste de procédures prenant en charge ces données.

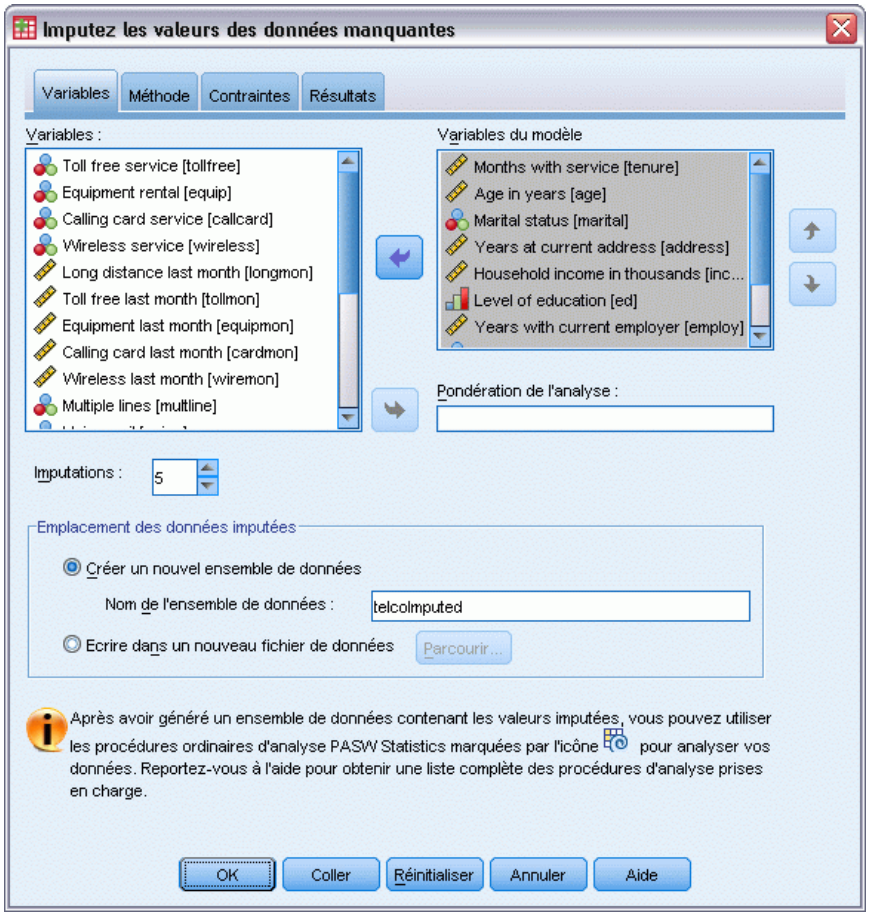
Exemple : Un fournisseur de services de télécommunication souhaite mieux comprendre les types d'utilisation des services dans sa base de données client. Il dispose de données complètes sur les services utilisés par les clients, mais les informations démographiques collectées par l'entreprise comportent certaines valeurs manquantes. De plus, ces valeurs ne sont pas manquantes de façon complètement aléatoire. Par conséquent, l'imputation multiple sera utilisée pour compléter l'ensemble de données. Pour plus d'informations, reportez-vous à la section Utilisation de l'imputation multiple pour compléter et analyser un ensemble de données dans le chapitre 5 sur p. 52.

Pour imputer les valeurs de données manquantes

A partir des menus, sélectionnez :

Analyse > Imputation multiple > Imputer les valeurs des données manquantes…

Figure 3-2
Imputer les valeurs de données manquantes Onglet Variables



- Sélectionner au moins deux variables dans le modèle d'imputation. La procédure impute des valeurs multiples pour les valeurs manquantes de ces variables.

- Spécifiez le nombre d'imputations à calculer. Par défaut, cette valeur est 5.

- Spécifiez un ensemble de données ou un fichier de données au format IBM® SPSS® Statistics dans lequel les données imputées devront être écrites.

  L'ensemble de données de sortie comprend les données d'observation initiales avec des données manquantes, ainsi qu'un ensemble d'observations avec des valeurs imputées pour chaque imputation. Par exemple, si l'ensemble de données initial comprend 100 observations et que vous ayez 5 imputations, l'ensemble de données de sortie comportera 600 observations. Toutes les variables dans l'ensemble de données d'entrée sont incluses dans l'ensemble de données de sortie. Les propriétés du dictionnaire (noms, étiquettes, etc.) des variables existantes sont copiées dans le nouvel ensemble de données. Le fichier contient également une nouvelle variable, *Imputation_*, une variable numérique qui indique l'imputation (0 pour les données d'origine, ou 1..*n* pour les observations ayant des valeurs imputées).

La procédure définit automatiquement la variable *Imputation_* comme variable de scission après la création de l'ensemble de données de sortie. Si des scissions sont actives lorsque la procédure est exécutée, l'ensemble de données de sortie comprend un ensemble d'imputations pour chaque combinaison de valeurs de variables de scission.

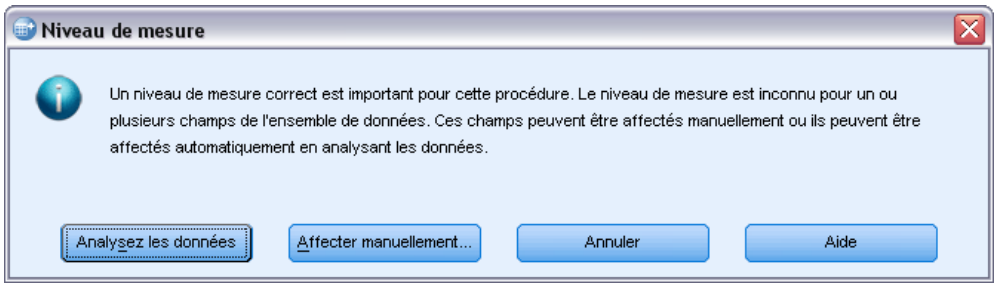
Paramètres facultatifs

Pondération d'analyse. Cette variable contient des pondérations (de régression ou d'échantillon) d'analyse. La procédure intègre des pondérations d'analyse en régression et des modèles de classification utilisés pour imputer les valeurs manquantes. Les pondérations d'analyse sont également utilisées dans les récapitulatifs de valeurs imputées ; par exemple, la moyenne, l'écart-type et l'erreur standard. Les observations ayant une pondération d'analyse négative ou nulle sont exclues.

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 3-3
Alerte du niveau de mesure

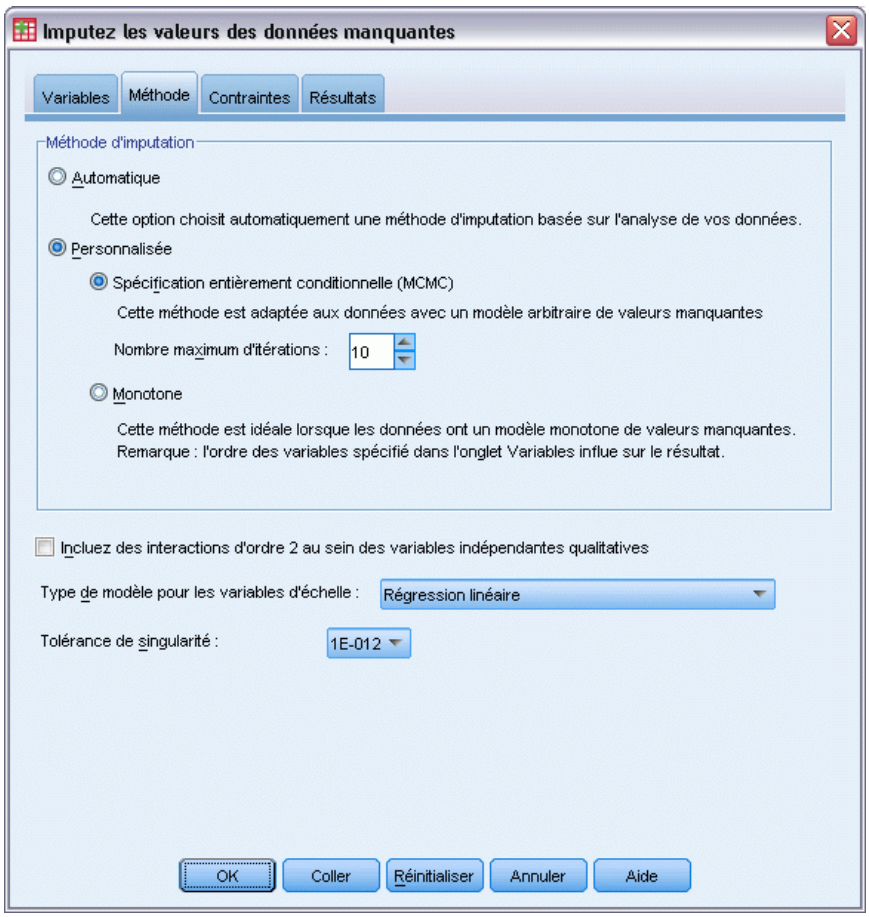


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Méthode

Figure 3-4
Imputer les valeurs de données manquantes Onglet Méthode



L'onglet Méthode spécifie de quelle manière les valeurs manquantes seront imputées, y compris les types des modèles utilisés. Les valeurs prédites sont codées par indicateurs (factices).

Méthode d'imputation. La méthode Automatique analyse les données et utilise la méthode monotone si les données présentent un modèle de valeurs manquantes monotone ; le reste du temps, la spécification entièrement conditionnelle est utilisée. Si vous êtes certain de la méthode à utiliser, vous pouvez la spécifier comme méthode personnalisée.

- **Spécification entièrement conditionnelle.** Il s'agit d'une méthode de Monte Carlo par chaînes de Markov (MCMC) itérative pouvant être utilisée lorsque le modèle de données manquantes est arbitraire (monotone ou non).

  Pour chaque itération et pour chaque variable dans l'odre spécifié par la liste de variables, la méthode de spécification entièrement conditionnelle (FCS) ajuste un modèle univarié (variable dépendante unique) en utilisant toutes les autres variables du modèle comme variables prédites, et impute ensuite les valeurs manquantes pour la variable à ajuster. Cette méthode se poursuit jusqu'à ce que le nombre maximal d'itérations soit atteint, et les valeurs imputées à l'itération maximale sont enregistrées dans l'ensemble de données imputé.

Nombre maximum d'itérations : Spécifie le nombre d'itérations, ou "d'étapes", utilisées par les chaînes de Markov dans la méthode FCS. Si la méthode FCS a été choisie automatiquement, elle utilise 10 itérations par défaut. Lorsque vous avez précisément choisi FCS, vous pouvez spécifier un nombre d'itérations personnalisé. Vous pourriez avoir à augmenter le nombre d'itérations si la chaîne de Markov n'a pas convergé. Dans l'onglet Résultats, vous pouvez enregistrer les données de l'historique des itérations FCS et les visualiser sous forme de diagramme pour évaluer la convergence.

- **Monotone**. Méthode non-itérative pouvant être utilisée uniquement lorsque les données présentent un modèle de valeurs manquantes monotone. Un modèle monotone existe lorsqu'il est possible d'ordonner les variables de façon à ce que, si une variable a une valeur non manquante, toutes les variables précédentes auront également des valeurs non manquantes. Lorsque vous la spécifiez comme une méthode Personnalisée, veillez à spécifier les variables de la liste dans un ordre faisant apparaître un modèle monotone.

  Pour chaque variable de l'ordre monotone, la méthode monotone ajuste un modèle univarié (variable dépendante unique) en utilisant toutes les variables précédentes comme variables prédites, et impute ensuite les valeurs manquantes pour la variable à ajuster. Ces valeurs imputées sont enregistrées dans l'ensemble de données imputé.

Inclure les interactions bidirectionnelles. Lorsque la méthode d'imputation est automatiquement choisie, le modèle d'imputation de chaque variable comprend un terme constant et des effets majeurs pour les variables prédites. Losqu'une méthode spécifique est choisie, vous pouvez, si vous le désirez, inclure toutes les interactions bidirectionnelles possibles parmi les variables prédites catégorielles.

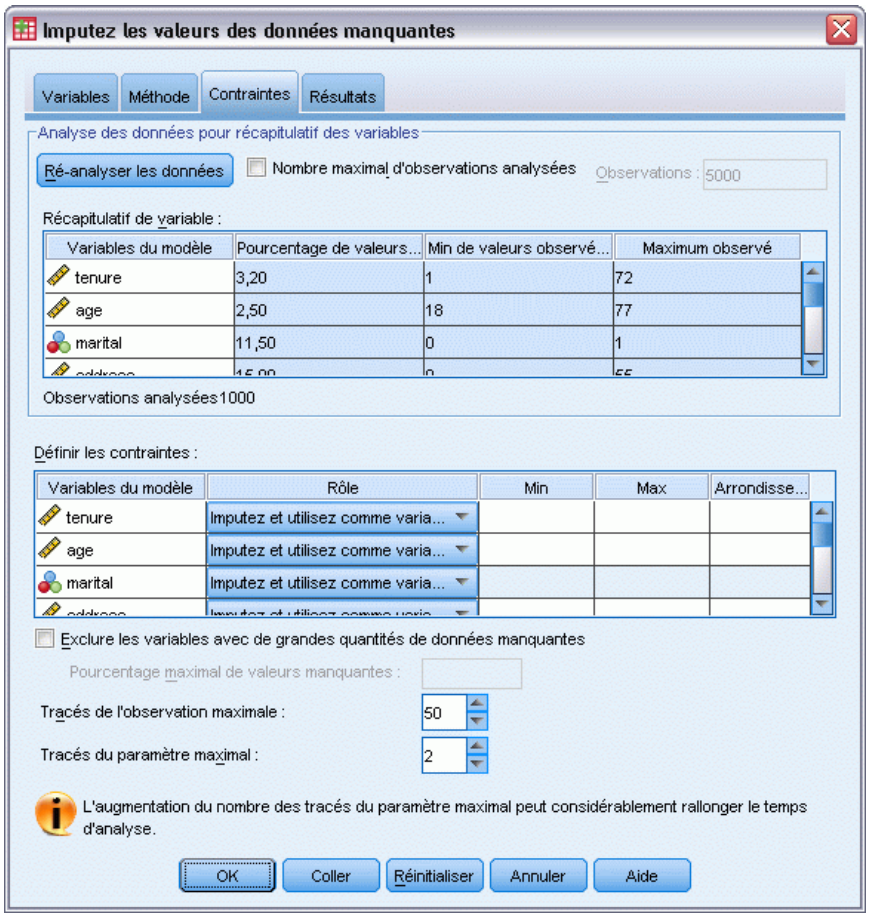
Type de modèle pour les variables d'échelle. Lorsque la méthode d'imputation est automatiquement sélectionnée, la régression linéaire est utilisée comme modèle univarié pour les variables d'échelle. Lorsqu'une méthode spécifique est choisie, vous pouvez également choisir l'égalisation par la moyenne prédictive (PMM) comme modèle pour les variables d'échelle. La méthode PMM est une variante de régression linéaire qui faire concorder les valeurs imputées par le modèle de régression et la valeur observée la plus proche.

La régression logistique est toujours utilisée comme modèle univarié pour les variables catégorielles. Indépendamment du type de modèle, les variables prédites qualitatives sont traitées à l'aide du codage par indicateurs (factice).

Tolérance singularité : Les matrices singulières (ou non inversables) comportent des colonnes linéairement dépendantes, ce qui peut provoquer de graves problèmes pour l'algorithme d'estimation. Même les matrices presque singulières peuvent générer des résultats médiocres. C'est pourquoi la procédure traite une matrice dont le déterminant est inférieur à la tolérance en tant que matrice singulière. Indiquez une valeur positive.

Contraintes

Figure 3-5
Imputer les valeurs de données manquantes Onglet Contraintes



L'onglet Contraintes vous permet de restreindre le rôle d'une variable pendant l'imputation et de restreindre la plage des valeurs imputées d'une variable d'échelle afin qu'elles soient plausibles. De plus, vous pouvez restreindre l'analyse aux variables avec moins d'un pourcentage maximal de valeurs manquantes.

Analyse des données pour le récapitulatif des variables. En cliquant sur Analyse des données , la liste affiche des variables d'analyse et le pourcentage observé manquant, minimum et maximum de chacune. Les récapitulatifs peuvent être basés sur toutes les observations ou limités à une analyse des *n* premières observations comme spécifié dans la zone de texte Observations. Pour mettre à jour les récapitulatifs de distribution, cliquez sur Réanalyser les données.

Définir les contraintes

- **Rôle.** Vous permet de personnaliser l'ensemble des variables à imputer et/ou à traiter comme variables prédites. Généralement, chaque variable d'analyse est considérée à la fois comme une variable dépendante et comme une variable prédite dans le modèle d'imputation. Le Rôle peut servir à désactiver l'imputation pour les variables que vous souhaitez Utiliser comme

variable prédite uniquement ou pour que des variables ne soient pas utilisées comme des valeurs prédites (Imputer uniquement) et obtenir ainsi des modèles plus compacts. C'est la seule contrainte qui peut être spécifiée pour les variables catégorielles, ou pour les variables qui sont uniquement utilisées comme valeurs prédites.

- **Min et Max.** Ces colonnes vous permettent de spécifier les valeurs imputées minimum et maximum autorisées pour les variables d'échelle. Si une valeur imputée dépasse cette plage, la procédure essaie une autre valeur jusqu'à ce qu'elle en trouve une qui soit dans la plage ou que le nombre maximum d'essais soit atteint (Consultez Essais maximum ci-dessous). Ces colonnes ne sont disponibles que si la régression linéaire est sélectionnée comme type de modèle de variable d'échelle dans l'onglet Méthode.
- **Arrondi.** Certaines variables peuvent être utilisées comme variables d'échelle, mais elles possèdent des valeurs par nature davantage restreintes. Par exemple, le nombre de personnes dans un ménage doit être un entier, et le montant dépensé lors d'un passage dans une épicerie ne peut contenir de centimes fractionnels. Cette colonne vous permet de spécifier la coupure la plus faible à accepter. Par exemple, pour obtenir des valeurs entières, vous devez spécifier 1 comme la coupure d'arrondissement et pour obtenir les valeurs arrondies au centime le plus proche, vous devez spécifier 0,01. Les valeurs sont généralement arrondies au multiple entier le plus proche de la coupure d'arrondissement. Le tableau suivant montre de quelle manière les valeurs arrondies agissent sur la valeur imputée de 6,64823 (avant arrondissement).

Coupure d'arrondissement	Valeur à laquelle 6,64832 est arrondie
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

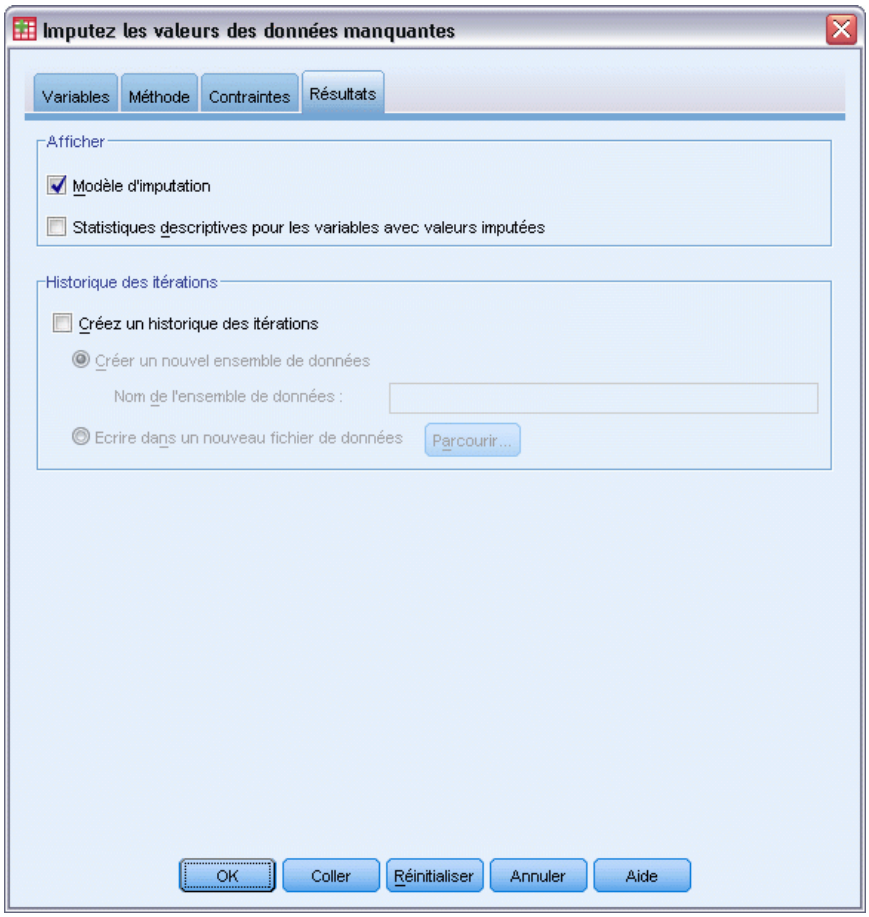
Excluent les variables avec de nombreuses données manquantes. Généralement, les variables d'analyse sont imputées et utilisées comme valeurs prédites sans tenir compte du nombre de leurs valeurs manquantes, tant qu'elles ont assez de données pour évaluer un modèle d'imputation. Vous pouvez choisir d'exclure des variables ayant un pourcentage élevé de valeurs manquantes. Par exemple, si vous spécifiez 50 comme Pourcentage maximum manquant, les variables d'analyse qui contiennent plus de 50% de valeurs manquantes ne sont pas imputées et ne sont pas non plus utilisées comme valeurs prédites dans les modèles d'imputation.

Essais maximum. Si des valeurs minimum ou maximum sont spécifiées pour les valeurs imputées des variables d'échelle (voir Min et Max ci-dessus), la procédure essaie de rechercher des valeurs jusqu'à ce qu'elle trouve un ensemble de valeurs dans les limites des plages spécifiées. Si un ensemble de valeurs n'est pas obtenu après avoir atteint le nombre d'essais par observation spécifié, la procédure essaie un autre ensemble de paramètres de modèle et répète la procédure d'essais d'observations. Une erreur se produit si un ensemble de valeurs dans la limite des plages n'est pas obtenu en respectant le nombre d'essais d'observations et de paramètres spécifié.

Veuillez noter que l'augmentation de ces valeurs peut augmenter la durée d'exécution. Si la procédure dure longtemps, ou n'est pas capable de trouver des essais appropriés, vérifiez les valeurs minimum et maximum spécifiées pour vous assurer qu'elles sont appropriées.

Résultats

Figure 3-6
Imputer les valeurs de données manquantes Onglet Résultats



Afficher : Affichage des commandes de sortie. Un récapitulatif général des imputations est toujours affiché et comprend des tableaux présentant les spécifications des imputations, les itérations (pour la méthode de spécification entièrement conditionnelle) des imputations, les variables dépendantes imputées, les variables dépendantes exclues de l'imputation et la séquence d'imputation. Si cette option est sélectionnée, les contraintes des variables d'analyse apparaissent également.

- **Modèle d'imputation.** Affiche le modèle d'imputation pour les variables dépendantes et pour les variables prédites et contient le type de modèle univarié, les effets de modèle et le nombre de valeurs imputées.
- **Statistiques descriptives :** Affiche les statistiques descriptives pour les variables dépendantes dont les valeurs sont imputées. Pour les variables d'échelle, les statistiques descriptives comprennent la moyenne, l'effectif, l'écart-type, le minimum et le maximum pour les données d'entrée d'origine (avant l'imputation), les valeurs imputées (par imputation) et les données complètes (à la fois les valeurs d'origine et imputées—par imputation). Pour les variables catégorielles, les statistiques descriptives comprennent l'effectif et le pourcentage par catégorie pour les données d'entrée d'origine (avant l'imputation), les valeurs imputées

(par imputation) et les données complètes (à la fois les valeurs d'origine et imputées—par imputation).

Historique des itérations. Lorsque la méthode d'imputation à spécification entièrement conditionnelle est utilisée, vous pouvez demander un ensemble de données contenant les données de l'historique des itérations pour l'imputation FCS. L'ensemble de données contient les moyennes et les écarts-types par itération et par imputation pour chaque variable d'échelle dépendante dont les valeurs sont imputées. Vous pouvez visualiser les données sous forme de graphique pour mieux évaluer la convergence du modèle. Pour plus d'informations, reportez-vous à la section Vérification de la convergence FCS dans le chapitre 5 sur p. 71.

Commande IMPUTATION MULTIPLE - Descriptives additionnelles

Le langage de syntaxe de commande vous permet aussi de :

- spécifier un sous-ensemble de variables dont les statistiques descriptives sont affichées (sous-commande`RECAPITULATIFSIMPUTATIONS`).
- Spécifier à la fois une analyse de modèles manquants et de l'imputation en n'exécutant la procédure qu'une seule fois.
- Spécifiez le nombre maximal de paramètres de modèle autorisé lors de l'imputation d'une variable (mot-clé`MAXMODELPARAM`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Utilisation des données à imputation multiple

Lorsqu'un ensemble de données à imputation multiple (IM) est créé, une variable appelée *Imputation_* avec une étiquette de variable *Nombre d'imputations* est ajoutée et l'ensemble de données est trié dans l'ordre croissant. Les observations de l'ensemble de données d'origine ont une valeur de 0. Les observations pour les valeurs imputées sont numérotées de 1 à *M*, où *M* est le nombre d'imputations.

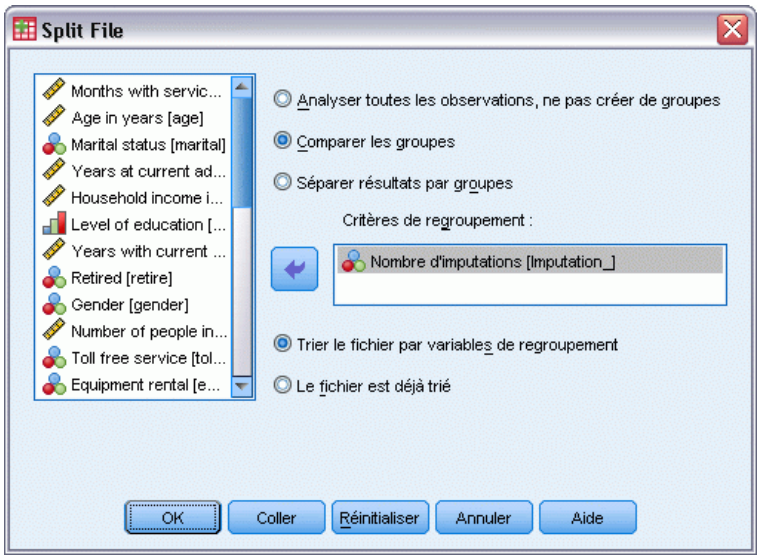
Lorsque vous ouvrez un ensemble de données, la présence de la variable *Imputation_* identifie l'ensemble de données comme un ensemble de données IM possible.

Activation d'un ensemble de données à imputation multiple pour l'analyse

L'ensemble de données doit être scindé à l'aide de l'option Comparer les groupes, avec *Imputation_* comme variable de regroupement, afin d'être traité comme un ensemble de données à imputation multiple lors des analyses. Vous pouvez également définir les scissions dans d'autres variables.

A partir des menus, sélectionnez :
Données > Scinder un fichier

Figure 3-7
Boîte de dialogue Scinder un fichier



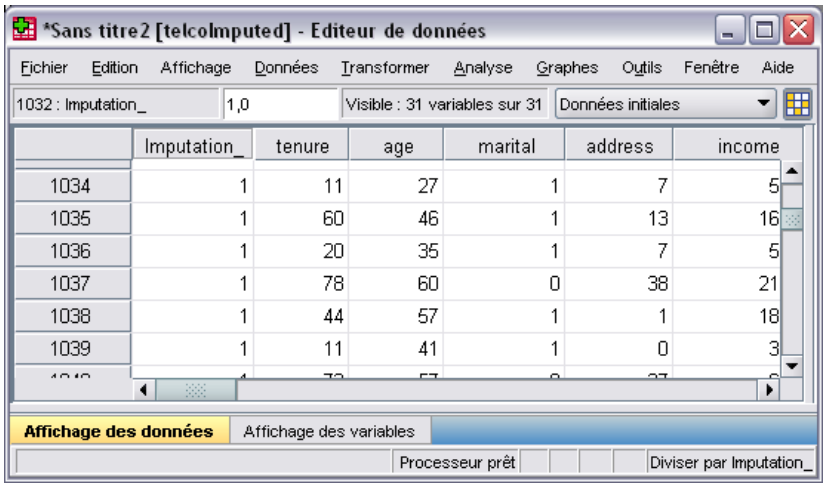
- Sélectionnez Comparer les groupes.
- Sélectionnez le *nombre d'imputations [Imputation_]* comme variable de regroupement des observations.

Egalement, lorsque vous activez le marquage (voir ci-dessous), le ficher est scindé par rapport au *nombre d'imputations [Imputation_]*.

Distinguer les valeurs imputées des valeurs observées

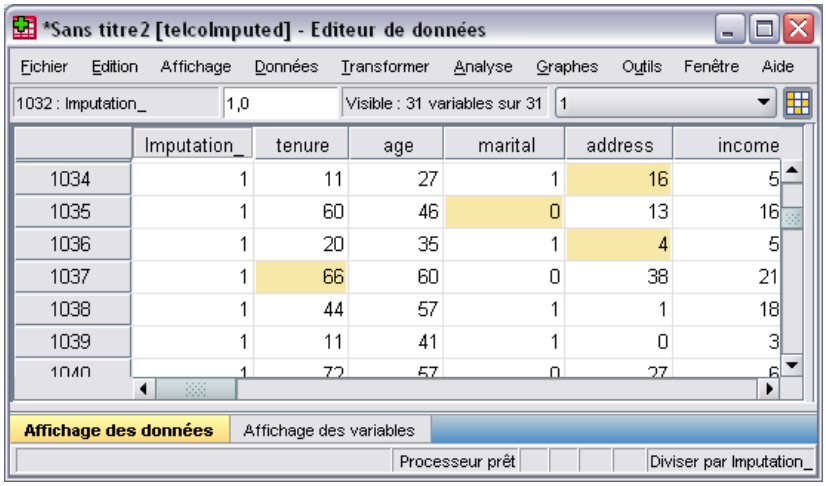
Vous pouvez distinguer les valeurs imputées des valeurs observées par la couleur d'arrière-plan des cellules, la police et l'écriture en gras (pour les valeurs imputées). Pour des détails avec marquage actif, consultez Options d'imputation multiple sur p. 35. Lorsque vous créez un nouvel ensemble de données dans la session actuelle avec l'option Imputer les valeurs manquantes, le marquage est activé par défaut. Lorsque vous ouvrez un fichier de données enregistré qui comprend des imputations, le marquage est désactivé.

Figure 3-8
L'éditeur de données avec marquage des imputations désactivé (OFF)



Pour activer le marquage, dans les menus de l'éditeur de données, choisissez :
Affichage > Marquer les données imputées...

Figure 3-9
L'éditeur de données avec marquage des imputations activé (ON)



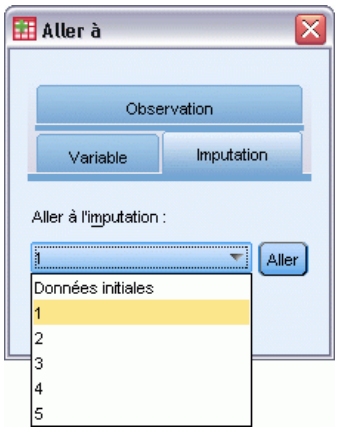
Vous pouvez également activer le marquage en cliquant sur le bouton d'activation du marquage des imputations sur le côté droit de la barre d'édition dans l'Affichage des données de l'éditeur de données.

Déplacement entre les imputations

- A partir des menus, sélectionnez :
  Edition > Aller à l'imputation...

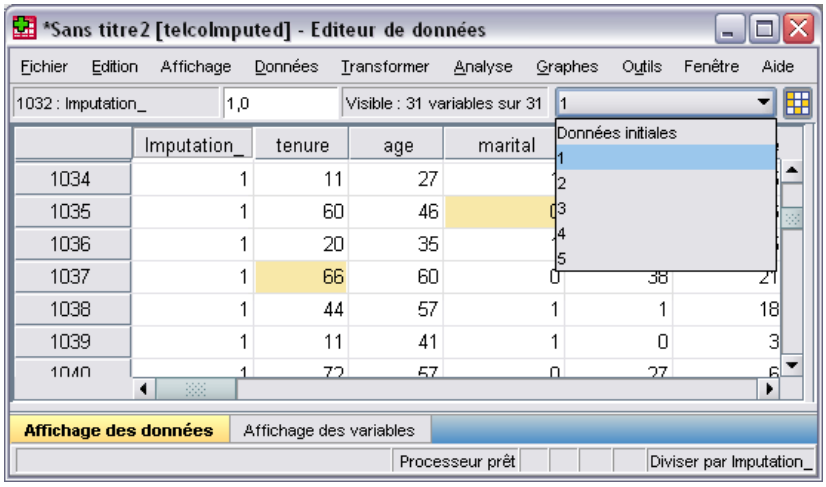
- Sélectionnez l'imputation (ou données d'origine) dans la liste déroulante proposée.

Figure 3-10
Boîte de dialogue Aller à



Vous pouvez également sélectionner l'imputation dans la liste déroulante de la barre d'édition dans l'Affichage des données de l'Éditeur de données.

Figure 3-11
L'éditeur de données avec marquage des imputations activé (ON)



La position relative des observations est conservée lors de la sélection des imputations. Par exemple, si l'ensemble de données initial contient 1000 observations, l'observation 1034, la 34ème observation de la première imputation, apparaît en haut de la grille. Si vous sélectionnez l'imputation 2 dans la liste déroulante, l'observation 2034, 34ème observation de l'imputation 2, apparaît en haut de la grille. Si vous sélectionnez Données d'origine dans la liste déroulante, l'observation 34 apparaît en haut de la grille. La position des colonnes est également conservée lorsque vous naviguez entre les imputations, pour une comparaison facile des valeurs entre les imputations.

Transformation et modification des valeurs imputées

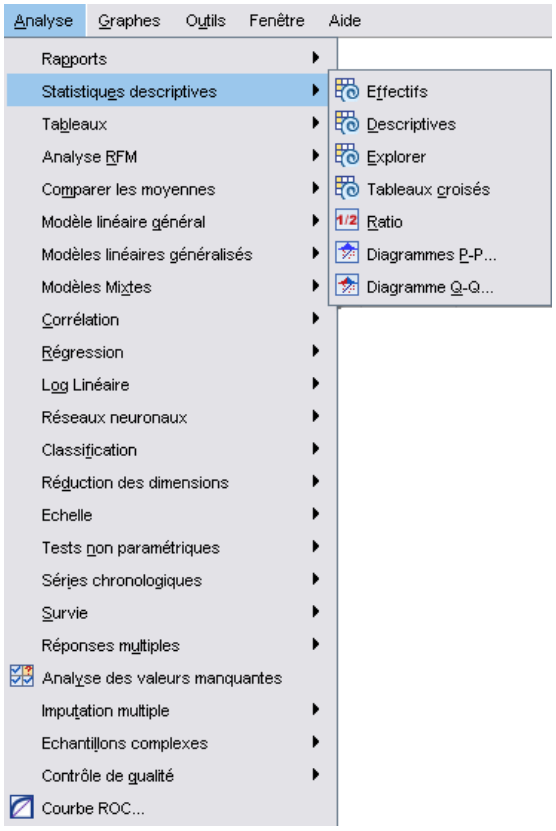
Parfois, vous aurez besoin d'effectuer des transformations sur les donnés imputées. Par exemple, vous pouvez décider de prendre le log de toutes les valeurs d'une variable de salaire et d'enregistrer le résultat dans une nouvelle variable. Une valeur calculée à l'aide des données imputées sera traitée comme imputée si elle diffère de la valeur calculée à l'aide des données d'origine.

Si vous modifiez une valeur imputée dans une cellule de l'éditeur de données, cette cellule sera traitée comme imputée. Nous vous déconseillons de modifier des valeurs imputées de cette façon.

Analyse de données à imputation multiple

De nombreuses procédures prennent en charge le regroupement de résultats d'une analyse d'ensembles de données à imputation multiple. Lorsque le marquage des imputations est activé, une icône spéciale apparaît à côté des procédures qui prennent en charge le regroupement. Dans le sous-menu Statistiques descriptives du menu Analyser par exemple, les procédures Effectifs, Descriptives, Explorer et Tableaux croisés prennent toutes en charge le regroupement, contrairement aux procédures Rapport, Diagrammes P-P et Diagrammes Q-Q .

Figure 3-12
Anlyser le menu avec marquage des imputations activé (ON)



Les tableaux de résultats et les modèles PMML peuvent être regroupés. Il n'existe pas de nouvelle procédure permettant de demander des résultats regroupés, mais un nouvel onglet de la boîte de dialogue Options vous permet de contrôler tous les résultats d'imputation multiple.

- **Regroupement des tableaux de résultats.** Par défaut, lorsque vous exécutez une procédure prise en charge dans un ensemble de données d'imputation multiple (IM), les résultats sont automatiquement produits pour chaque imputation, pour les données d'origine (non imputées) et pour les résultats regroupés (finaux) qui prennent en compte les variations entre les imputations. Les statisques qui sont combinées varient selon la procédure.
- **Regroupement de PMML.** Vous pouvez également obtenir des PMML regroupés à partir des procédures prises en charge qui exportent les PMML. Le PMML regroupé est demandé de la même façon que le PMML non regroupé (qu'il remplace lorsqu'il est enregistré).

Les procédures non prises en charge ne produisent ni résultats regroupés ni fichiers PMML regroupés.

Niveaux de combinaison

Les résultats sont regroupés à l'un des deux niveaux suivants :

- Combinaison**Naïve.** Seul le paramètre regroupé est disponible.
- **Combinaison univariée.** Le paramètre regroupé, son erreur standard, sa statistique de test et ses degrés réels de liberté, la valeur *p*, l'intervalle de confiance et les diagnostiques de regroupements (fraction des informations manquantes, efficacité relative, augmentation relative de la variance) sont affichés lorsqu'ils sont disponibles.

Les coefficients (régression et corrélation), les moyennes (et différences moyennes) et les effectifs sont généralement combinés. Lorsque l'erreur standard d'une statistique est disponible, le regroupement univarié est alors utilisé. Autrement, c'est le regroupement simpliste qui est utilisé.

Procédures prenant en charge le regroupement

Les procédures suivantes prennent en charge les ensembles de données IM, avec le niveau de regroupement spécifié pour chaque partie des résultats.

Effectifs

- Le tableau Statistiques prend en charge les Moyennes en regroupement univarié (si la moyenne E.S. est également requise), ainsi que N Valide et N manquant pour le regroupement Naïve.
- Le tableau Effectifs prend en charge les effectifs en regroupement Naïve.

Descriptives

- Le tableau Statistiques prend en charge les Moyennes en regroupement univarié (si la moyenne E.S. est également requise), ainsi que N pour le regroupement Naïve.

Tableaux croisés

- Le tableau croisé prend en charge les effectifs en regroupement Naïve.

Moyennes

- Le tableau Rapport prend en charge la moyenne en regroupement univarié (si la moyenne E.S. est également requise), ainsi que N pour le regroupement Naïve.

Test T pour échantillon unique

- Le tableau Statistiques prend en charge la moyenne en regroupement univarié et N en regroupement Naïve.
- Le tableau Test prend en charge la différence moyenne en regroupement Naïve.

Test T pour échantillons indépendants

- Le tableau Statistiques de groupes prend en charge les moyennes en regroupement univarié et N en regroupement Naïve.
- Le tableau Test prend en charge la différence moyenne en regroupement univarié.

Test T pour échantillons appariés

- Le tableau Statistiques prend en charge les moyennes en regroupement univarié et N en regroupement Naïve.
- Le tableau Corrélations prend en charge les corrélations et N en regroupement Naïve.
- Le tableau Test prend en charge la moyenne en regroupement univarié.

ANOVA à 1 facteur

- Le tableau Statistiques descriptives prend en charge la moyenne en regroupement univarié et N en regroupement Naïve.
- Le tableau Tests de contraste prend en charge la valeur du contraste en regroupement univarié.

GML univarié, GML multivarié et GML répété

- Le tableau Facteurs inter-sujets prend en charge N en regroupement Naïve.
- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement Naïve.
- La tableau Estimations de paramètre prend en charge le coefficient, B, en regroupement univarié.
- Les moyennes marginales estimées : Le tableau Estimations prend en charge la moyenne en regroupement univarié.
- Les moyennes marginales estimées : Le tableau Comparaisons par paire prend en charge la différence moyenne en regroupement univarié.

Modèles mixtes linéaires

- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement Naïve.
- Le tableau Estimations des effets fixes prend en charge les estimations en regroupement univarié.
- Le tableau Estimations des paramètres de covariance prend en charge les estimations en regroupement univarié.

- Les moyennes marginales estimées : Le tableau Estimations prend en charge la moyenne en regroupement univarié.
- Les moyennes marginales estimées : Le tableau Comparaisons par paire prend en charge la différence moyenne en regroupement univarié.

Modèles linéaires généralisés et équations d'estimation généralisées. Ces procédures prennent en charge le PMML regroupé.

- Le tableau Informations sur les variables catégorielles prend en charge N et les pourcentages en regroupement Naïve.
- Le tableau Informations sur les variables continues prend en charge N et les pourcentages en regroupement Naïve.
- La tableau Estimations de paramètre prend en charge le coefficient, B, en regroupement univarié.
- Les moyennes marginales estimées : Le tableau Coefficients d'estimation prend en charge la moyenne en regroupement Naïve.
- Les moyennes marginales estimées : Le tableau Estimations prend en charge la moyenne en regroupement univarié.
- Les moyennes marginales estimées : Le tableau Comparaisons par paire prend en charge la différence moyenne en regroupement univarié.

Corrélations bivariées

- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement Naïve.
- Le tableau Corrélations prend en charge les corrélations et N en regroupement Naïve.

Corrélations partielles

- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement Naïve.
- Le tableau Corrélations prend en charge les corrélations en regroupement Naïve.

Régression linéaire. Cette procédure prend en charge le PMML regroupé.

- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement Naïve.
- Le tableau Corrélations prend en charge les corrélations et N en regroupement Naïve.
- Le tableau Coefficients prend en charge B en regroupement univarié et les corrélations en regroupement Naïve.
- Le tableau Coefficients de corrélation prend en charge les corrélations en regroupement Naïve.
- Le tableau Statistiques résiduelles prend en charge la moyenne et N en regroupement Naïve.

Régression logistique binaire. Cette procédure prend en charge le PMML regroupé.

- Le tableau Variables dans l'équation prend en charge B en regroupement univarié.

Régression logistique multinomiale. Cette procédure prend en charge le PMML regroupé.

- La tableau Estimations de paramètre prend en charge le coefficient, B, en regroupement univarié.

Régression ordinale

- La tableau Estimations de paramètre prend en charge le coefficient, B, en regroupement univarié.

Analyse discriminante. Cette procédure prend en charge le modèle XML regroupé.

- Le tableau Statistiques de groupes prend en charge la moyenne et N Valide en regroupement Naïve.
- Le tableau Matrices intra-classes globales prend en charge les corrélations en regroupement Naïve.
- La tableau Coefficients de fonction discriminante canonique prend en charge les coefficients non standardisés en regroupement Naïve.
- La tableau Fonctions aux barycentres des groupes prend en charge les coefficients non standardisés en regroupement Naïve.
- Le tableau Coefficients de fonction de classification prend en charge les coefficients en regroupement Naïve.

Test du Khi-deux

- Le tableau Descriptives prend en charge la moyenne et N en regroupement Naïve.
- Le tableau Effectifs prend en charge N observé en regroupement Naïve.

Test binomial

- Le tableau Descriptives prend en charge les moyennes et N en regroupement Naïve.
- Le tableau Test prend en charge N, la proportion observée et le test de proportion en regroupement Naïve.

Suites en séquences

- Le tableau Descriptives prend en charge les moyennes et N en regroupement Naïve.

Test Kolmogorov-Smirnov pour un échantillon

- Le tableau Descriptives prend en charge les moyennes et N en regroupement Naïve.

Tests pour deux échantillons indépendants

- Le tableau Rangs prend en charge le rang moyen et N en regroupement Naïve.
- Le tableau Effectifs prend en charge N en regroupement Naïve.

Tests pour plusieurs échantillons indépendants

- Le tableau Rangs prend en charge le rang moyen et N en regroupement Naïve.
- Le tableau Effectifs prend en charge les effectifs en regroupement Naïve.

Tests pour deux échantillons liés

- Le tableau Rangs prend en charge le rang moyen et N en regroupement Naïve.
- Le tableau Effectifs prend en charge N en regroupement Naïve.

Tests pour plusieurs échantillons liés

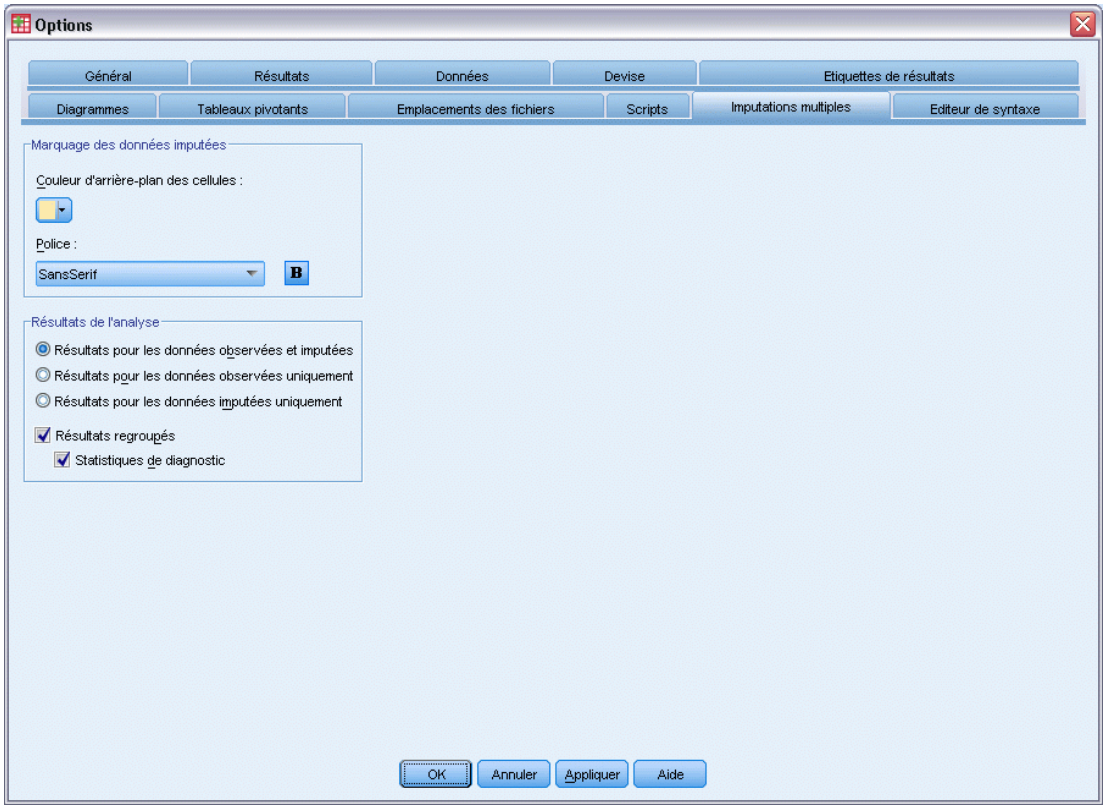
- Le tableau Rangs prend en charge le rang moyen en regroupement Naïve.

Régression de Cox. Cette procédure prend en charge le PMML regroupé.

- Le tableau Variables dans l'équation prend en charge B en regroupement univarié.
- Le tableau Moyennes des covariables prend en charge la moyenne en regroupement Naïve.

Options d'imputation multiple

Figure 3-13
Boîte de dialogue Options : Onglet Imputations multiples



L'onglet Imputations multiples contrôle deux sortes de préférences asssociées aux imputations multiples :

L'apparence des données imputées. Par défaut, les cellules contenant des données imputées auront un arrière-plan d'une autre couleur que celui des cellules contenant des données non-imputées. Cette différence d'apparence des données imputées devrait faciliter la navigation dans les ensembles de données et la recherche de ces cellules. Vous pouvez modifier la couleur d'arrière-plan par défaut des cellules, la police et afficher les données imputées en gras.

Résultats d'analyse. Ce groupe contrôle le type de résultats du Viewer produits lorsqu'un ensemble de données à imputation multiple est analysé. Par défaut, les résultats seront produits pour l'ensemble de données d'origine (pré-imputation) et pour chacun des ensembles de données imputés. De plus, pour ce genre de procédures qui prennent en charge le regroupement de données imputées, des résultats combinés finaux seront générés. Lorsqu'un regroupement univarié sera effectué, les diagnostiques de regroupement seront également affichés. Mais vous pouvez supprimer tous les résultats que vous ne désirez pas voir.

Pour définir les options d'imputation multiple

A partir du menu, sélectionnez :
Affichage > Options

Cliquez sur l'onglet Imputation multiple.

Partie II: Exemples

Chapitre
4

Analyse des valeurs manquantes

Description du modèle des données manquantes

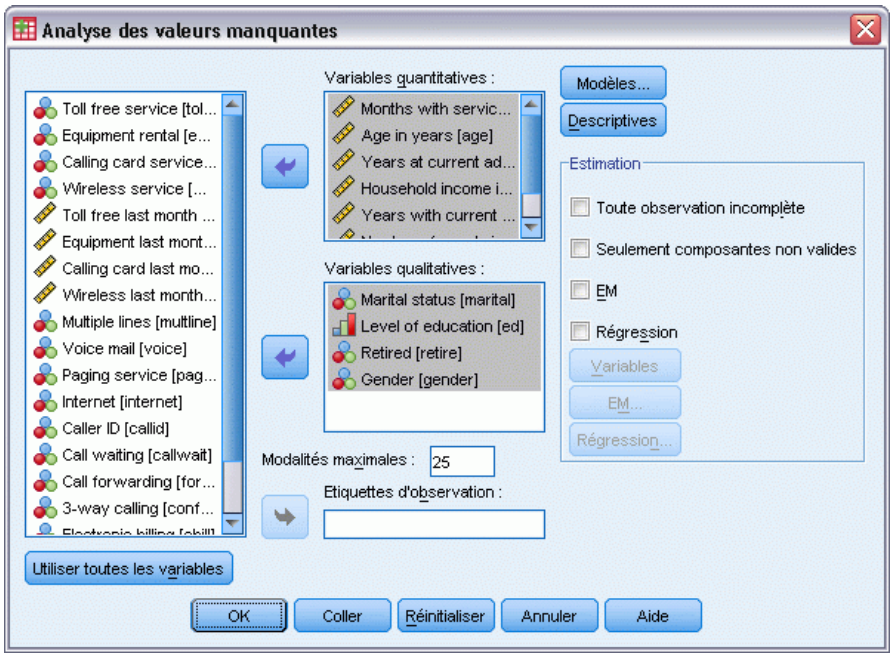
Un fournisseur de services de télécommunication souhaite mieux comprendre les types d'utilisation des services dans sa base de données client. La société souhaite s'assurer que les données sont des valeurs manquantes complètement aléatoires avant d'exécuter d'autres analyses.

Un échantillon aléatoire issu de la base de données client figure dans le fichier *telco_missing.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Missing Values 19*.

Exécution de l'analyse pour afficher les statistiques descriptives

- Pour exécuter l'analyse des valeurs manquantes, sélectionnez dans les menus :
  Analyse > Analyse des valeurs manquantes

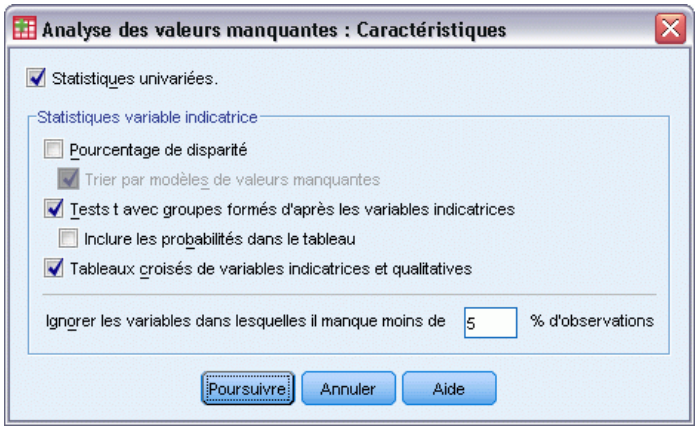
Figure 4-1
Boîte de dialogue Analyse des valeurs manquantes



- Sélectionnez *Marital status [marital]*, *Level of education [ed]*, *Retired [retire]* et *Gender* comme variables qualitatives.
- Sélectionnez *Months with service [tenure]* et *Number of people in household [reside]* comme variables (d'échelle) quantitatives.

  A ce stade, vous pourriez exécuter la procédure et obtenir des statistiques univariées, mais nous allons sélectionner des statistiques descriptives supplémentaires.
- Cliquez sur Descriptives.

Figure 4-2
Analyse des valeurs manquantes : Boîte de dialogue Descriptives



Dans la boîte de dialogue Descriptives, vous pouvez indiquer les statistiques descriptives à afficher dans le résultat. Les statistiques univariées par défaut vous permettent de déterminer l'étendue générale des données manquantes, mais les statistiques de variable indicatrice fournissent davantage d'informations sur l'impact du modèle de données manquantes d'une variable sur les valeurs d'une autre variable.

- Sélectionnez t tests avec groupes formés d'après les variables d'indication.
- Sélectionnez Tableaux croisés de variables indicatrices et qualitatives.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue principale Analyse des valeurs manquantes, cliquez sur OK.

Evaluation des statistiques descriptives

Dans cet exemple, les résultats comprennent les éléments suivants :

- Statistiques univariées.
- Tableau de tests *t* des variances séparées, y compris les moyennes de sous-groupes si une autre variable est présente ou manquante
- Tableaux pour chaque variable qualitative indiquant les effectifs des données manquantes pour chaque modalité par variable (d'échelle) quantitative

Figure 4-3
Tableau des statistiques univariées

	N	Moyenne	Ecart-type	Manquante		Nombre d'extrema[a]	
				Effectif	Pourcentage	Basse	Haute
MonthsWithService	968	35,56	21,268	32	3,2	0	0
Age	975	41,75	12,573	25	2,5	0	0
YearsAtAddress	850	11,47	9,965	150	15,0	0	9
Income	821	71,1462	83,14424	179	17,9	0	71
YearsWithEmployer	904	11,00	10,113	96	9,6	0	15
PeopleInHousehold	966	2,32	1,431	34	3,4	0	33
MaritalStatus	885			115	11,5		
EducationalLevel	965			35	3,5		
RetirementStatus	916			84	8,4		
Gender	958			42	4,2		

a. Nombre d'observations hors de l'intervalle (Q1 - 1,5*IQR, Q3 + 1,5*IQR).

Les statistiques univariées vous donnent un premier aperçu, variable par variable, de l'étendue des données manquantes. Le nombre de valeurs non manquantes pour chaque variable apparaît dans la colonne *N*, tandis que le nombre de valeurs manquantes figure dans la colonne *Nombre valeurs manquantes*. La colonne *Pourcentage valeurs manquantes* affiche le pourcentage d'observations comportant des valeurs manquantes et permet de comparer l'étendue des données manquantes parmi les variables. La variable *revenu (Revenu du ménage en milliers)* présente la plus forte proportion d'observations avec valeurs manquantes (17,9 %), tandis que la variable *âge (Âge en années)* affiche la plus faible (2.5%). La variable *revenu* présente le nombre le plus élevé de valeurs extrêmes.

Figure 4-4
Tableau des tests des variances séparées

		MonthsWithServic	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
address	t	,4	,3	.	3,5	1.4	1.0
	df	202,2	192,5	.	313,6	191.1	199.5
	# non manquantes	819	832	850	693	766	824
	#manquantes	149	143	0	128	138	142
	Moyenne (non manquantes)	35,68	41,79	11.47	74,0779	11.20	2.34
	Moyenne (manquantes)	34,91	41,49	.	55,2734	9.86	2.21
income	t	-5,0	-8,3	-3,9	.	-5.9	3.6
	df	249,5	222,8	191,1	.	203.3	315.2
	# non manquantes	793	801	693	821	741	792
	#manquantes	175	174	157	0	163	174
	Moyenne (non manquantes)	33,93	40,01	10,67	71,1462	9.91	2.39
	Moyenne (manquantes)	42,97	49,73	14,97	.	15.93	2.02
employ	t	-1,0	-,4	-,7	,5	.	-.3
	df	110,5	110,2	97,6	114,9	.	110,9
	# non manquantes	877	881	766	741	904	874
	#manquantes	91	94	84	80	0	92
	Moyenne (non manquantes)	35,34	41,69	11,37	71,4953	11,00	2,31
	Moyenne (manquantes)	37,70	42,27	12,32	67,9125	.	2,37
marital	t	,0	1,8	1,2	-,8	.9	-2.2
	df	148,1	149,5	138,8	121,2	128.3	134,2
	# non manquantes	856	862	748	728	805	857
	#manquantes	112	113	102	93	99	109
	Moyenne (non manquantes)	35,56	42,00	11,61	70,3887	11,10	2,28
	Moyenne (manquantes)	35,57	39,85	10,43	77,0753	10,17	2,61
retire	t	-,6	-,4	-.4	,3	.	.2
	df	95.4	94.4	84.0	93.2	.	99.0
	# non manquantes	888	893	777	751	904	885
	#manquantes	80	82	73	70	0	81
	Moyenne (non manquantes)	35,44	41,70	11,42	71,3356	11,00	2,32
	Moyenne (manquantes)	36,89	42,29	11,96	69,1143	.	2,30

Le tableau des tests *t* des variances séparées permet d'identifier les variables dont le modèle de valeurs manquantes peut influencer les variables (d'échelle) quantitatives. Le test *t* est calculé à l'aide d'une variable indicatrice qui indique si une variable est présente ou manquante pour une observation individuelle. Les moyennes de sous-groupes pour la variable indicatrice sont également mises en tableau. Une variable indicatrice n'est créée que si une variable possède des valeurs manquantes dans au moins 5 % des observations.

Il semble que les répondants plus âgés soient moins disposés à indiquer leurs niveaux de revenu. Lorsque la variable *revenu* est manquante, l'*âge* moyen est 49,73, contre 40,01 lorsque la variable *revenu* est non manquante. En effet, le descriptive manquant de la variable *revenu* semble avoir un impact sur les moyennes de plusieurs variables (d'échelle) quantitatives. Ceci est un signe que les données ne sont pas nécessairement des valeurs manquantes complètement aléatoires.

Figure 4-5
Tableau croisé de la variable Situation familiale [marital]

			Total	Unmarried	Married	Manquantes
						Manquantes système
YearsAtAddress	Non manquantes	Effectif	850	390	358	102
		Pourcentage	85,0	85,5	83,4	88,7
	Manquantes	% manquantes système	15,0	14,5	16,6	11,3
Income	Non manquantes	Effectif	821	380	348	93
		Pourcentage	82,1	83,3	81,1	80,9
	Manquantes	% manquantes système	17,9	16,7	18,9	19,1
YearsWithEmployer	Non manquantes	Effectif	904	418	387	99
		Pourcentage	90,4	91,7	90,2	86,1
	Manquantes	% manquantes système	9,6	8,3	9,8	13,9
RetirementStatus	Non manquantes	Effectif	916	423	392	101
		Pourcentage	91,6	92,8	91,4	87,8
	Manquantes	% manquantes système	8,4	7,2	8,6	12,2

Les variables indicatrices ayant moins de 5% de manquantes ne sont pas affichées.

Les tableaux croisés des variables qualitatives par rapport aux variables indicatrices présentent des informations similaires à celles du tableau des tests *t* des variances séparées. Des variables indicatrices sont de nouveau créées. Néanmoins, elles serviront cette fois à calculer les effectifs dans chaque modalité pour chaque variable qualitative. Les valeurs vous permettent de déterminer s'il existe des différences dans les valeurs manquantes parmi les modalités.

Le tableau concernant *marital (Situation familiale)* indique que le nombre de valeurs manquantes dans les variables indicatrices ne semble pas varier beaucoup entre les modalités de *marital*. Le fait qu'une personne soit mariée ou célibataire ne semble pas avoir d'incidence sur l'existence de données manquantes pour les variables (d'échelle) quantitatives. Par exemple, les personnes célibataires ont renseigné la variable *adresse (Nb d'années à la même adresse)* dans 85,5 % des cas, contre 83,4 % pour les personnes mariées. La différence est minime et vraisemblablement due au hasard.

Figure 4-6
Tableau croisé de Niveau d'éducation [ed]

			Catégories						
			Total	Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	Manquantes
									Manquantes système
YearsAtAddress	Non manquantes	Effectif	850	163	240	175	186	56	30
		Pourcentage	85,0	83,2	85,7	88,4	81,9	87,5	85,7
	Manquantes	% manquantes système	15,0	16,8	14,3	11,6	18,1	12,5	14,3
Income	Non manquantes	Effectif	821	155	229	165	193	50	29
		Pourcentage	82,1	79,1	81,8	83,3	85,0	78,1	82,9
	Manquantes	% manquantes système	17,9	20,9	18,2	16,7	15,0	21,9	17,1
YearsWithEmployer	Non manquantes	Effectif	904	178	254	178	204	60	30
		Pourcentage	90,4	90,8	90,7	89,9	89,9	93,8	85,7
	Manquantes	% manquantes système	9,6	9,2	9,3	10,1	10,1	6,2	14,3
MaritalStatus	Non manquantes	Effectif	885	193	278	148	184	52	30
		Pourcentage	88,5	98,5	99,3	74,7	81,1	81,2	85,7
	Manquantes	% manquantes système	11,5	1,5	,7	25,3	18,9	18,8	14,3
RetirementStatus	Non manquantes	Effectif	916	180	259	180	207	60	30
		Pourcentage	91,6	91,8	92,5	90,9	91,2	93,8	85,7
	Manquantes	% manquantes système	8,4	8,2	7,5	9,1	8,8	6,2	14,3

Maintenant, observons le tableau croisé concernant *ed (Niveau d'éducation)*. Si un répondant a poursuivi des études supérieures, une réponse pour la situation familiale est davantage susceptible d'être manquante. Au moins 98,5 % des répondants n'ayant pas poursuivi des études supérieures ont indiqué leur situation familiale. A l'opposé, seuls 81,1 % de ceux titulaires d'un diplôme universitaire ont indiqué leur situation familiale. Le nombre est encore moins élevé pour ceux ayant poursuivi des études supérieures, mais qui ne sont titulaires d'aucun diplôme universitaire.

Figure 4-7
Tableau croisé de Retraité [retire]

			Total	No	Yes	Manquantes
						Manquantes système
YearsAtAddress	Non manquantes	Effectif	850	744	33	73
		Pourcentage	85,0	85,0	80,5	86,9
	Manquantes	% manquantes système	15,0	15,0	19,5	13,1
Income	Non manquantes	Effectif	821	732	19	70
		Pourcentage	82,1	83,7	46,3	83,3
	Manquantes	% manquantes système	17,9	16,3	53,7	16,7
YearsWithEmployer	Non manquantes	Effectif	904	864	40	0
		Pourcentage	90,4	98,7	97,6	,0
	Manquantes	% manquantes système	9,6	1,3	2,4	100,0
MaritalStatus	Non manquantes	Effectif	885	777	38	70
		Pourcentage	88,5	88,8	92,7	83,3
	Manquantes	% manquantes système	11,5	11,2	7,3	16,7

Les variables indicatrices ayant moins de 5% de manquantes ne sont pas affichées.

Une différence plus marquée apparaît dans *retraite (retire)*. Les personnes à la retraite sont beaucoup moins susceptibles d'indiquer leur revenu que celles en activité. Seuls 46,3 % des clients à la retraite ont indiqué le niveau de revenu, contre 83,7 % de ceux en activité.

Figure 4-8
Tableau croisé pour Sexe [gender]

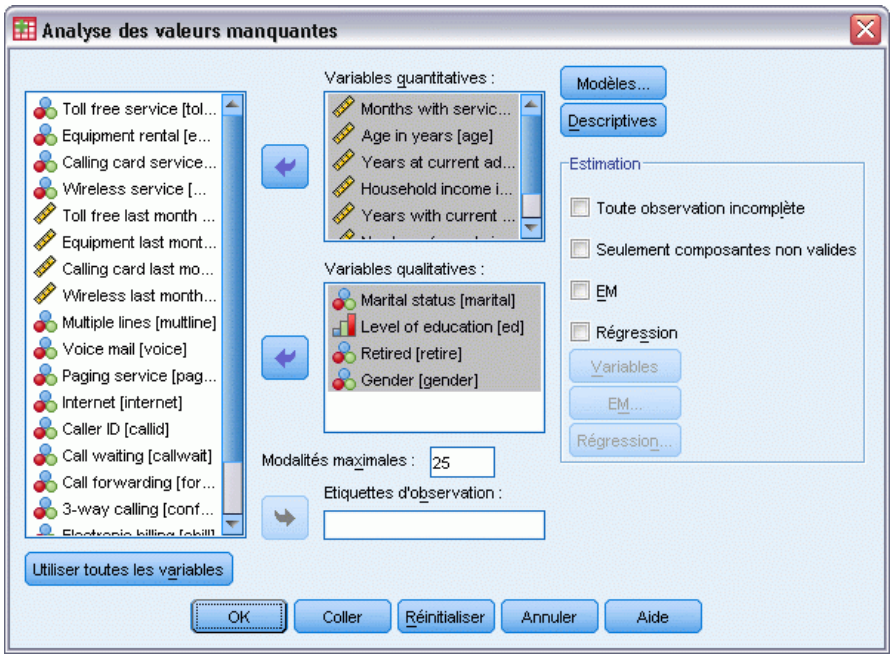
			Total	Male	Female	Manquantes
						Manquantes système
YearsAtAddress	Non manquantes	Effectif	850	363	456	31
		Pourcentage	85,0	78,6	91,9	73,8
	Manquantes	% manquantes système	15,0	21,4	8,1	26,2
Income	Non manquantes	Effectif	821	381	406	34
		Pourcentage	82,1	82,5	81,9	81,0
	Manquantes	% manquantes système	17,9	17,5	18,1	19,0
YearsWithEmployer	Non manquantes	Effectif	904	412	457	35
		Pourcentage	90,4	89,2	92,1	83,3
	Manquantes	% manquantes système	9,6	10,8	7,9	16,7
MaritalStatus	Non manquantes	Effectif	885	400	445	40
		Pourcentage	88,5	86,6	89,7	95,2
	Manquantes	% manquantes système	11,5	13,4	10,3	4,8
RetirementStatus	Non manquantes	Effectif	916	420	461	35
		Pourcentage	91,6	90,9	92,9	83,3
	Manquantes	% manquantes système	8,4	9,1	7,1	16,7

Une autre différence apparaît pour *sexe (Gender)*. Les informations d'adresse sont plus souvent manquantes pour les individus de sexe masculin que pour les individus de sexe féminin. Bien que ces différences puissent être dues au hasard, cela semble peu probable. Les données ne semblent pas correspondre à des valeurs manquantes complètement aléatoires.

Nous allons examiner les modèles de données manquantes afin d'en savoir plus.

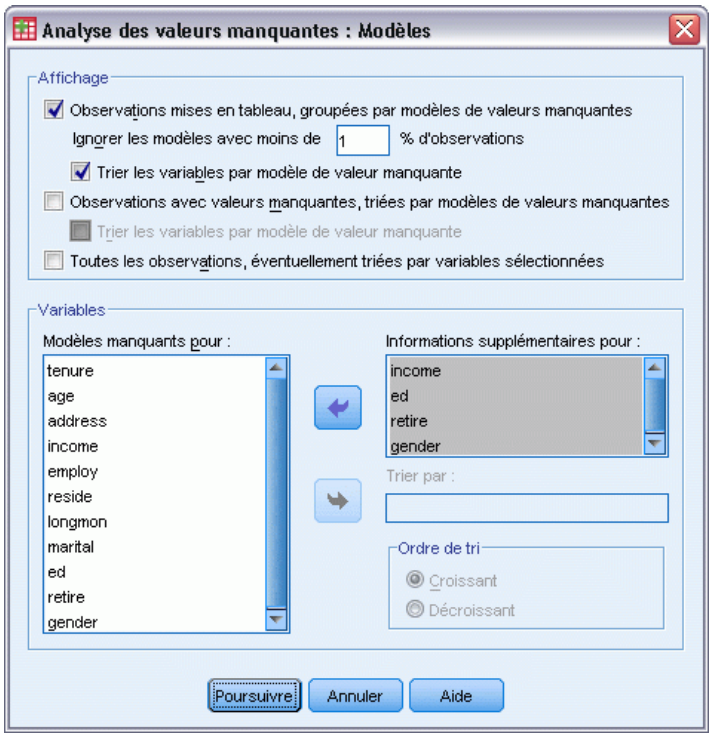
Réexécution de l'analyse pour afficher les modèles

Figure 4-9
Boîte de dialogue Analyse des valeurs manquantes



- Rappeler la boîte de dialogue Descriptives de l'analyse des valeurs manquantes. La boîte de dialogue conserve la variable utilisée dans l'analyse précédente. Ne la modifiez pas.
- Cliquez sur Modèles.

Figure 4-10
Boîte de dialogue Modèles d'analyses des valeurs manquantes



La boîte de dialogue Modèles vous permet de sélectionner différents tableaux de modèles. Nous allons afficher les modèles mis en tableau en les regroupant par modèle de valeurs manquantes. Etant donné que les modèles manquants dans *ed (Niveau d'éducation)*, *retire (Retraite)* et *gender (Sexe)* semblent avoir influencé les données, nous allons afficher des informations supplémentaires pour ces variables. Nous allons également inclure des informations supplémentaires pour *revenu (Revenu du ménage en milliers)*, en raison de son nombre élevé de valeurs manquantes.

- Sélectionnez Observations mises en tableau, groupées par modèles de valeurs manquantes.
- Sélectionnez *revenu*, *ed*, *retire* et *gender*, puis ajoutez-les à la liste Informations supplémentaires pour.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue principale Analyse des valeurs manquantes, cliquez sur OK.

Evaluation du tableau de modèles

Figure 4-11
Tableau des modèles mis en tableau

Nombre d'observations	Types de valeurs manquantes[a]										Complet si ...[b]	Income[c]	EducationalLevel[d]					Retirement Status[d]		Gender[d]	
	Age	PeopleInHousehold	MonthsWithService	EducationalLevel	Gender	RetirementStatus	YearsWithEmployer	MaritalStatus	YearsAtAddress	Income			Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	No	Yes	Male	Female
475											475	76,5853	99	157	87	101	31	463	12	201	274
109										X	584	.	27	35	19	17	11	95	14	47	62
16									X	X	687	.	5	9	0	1	1	12	4	12	4
87									X		562	54,4368	21	27	9	24	6	85	2	66	21
13		X									488	56,0000	4	3	2	3	1	13	0	4	9
60								X			535	77,2167	1	2	27	24	6	59	1	35	25
16				X							491	47,8125	0	0	0	0	0	16	0	6	10
17			X								492	76,2353	2	7	3	4	1	17	0	7	10
18					X						493	54,1111	3	7	4	4	0	17	1	0	0
16								X		X	660	.	0	0	7	8	1	14	2	6	10
37						X	X				520	59,4595	9	14	5	8	1	0	0	15	22

Les types ayant moins de 1% d'observations (10 ou moins) ne sont pas affichés.

a. Les variables sont triées en fonction des de valeurs manquantes.

b. Nombre d'observations complètes si les variables manquantes de ce type (identifiées par un X) ne sont pas utilisées.

c. Moyennes pour chaque type unique

d. Répartition des fréquences pour chaque type unique

Le tableau des modèles mis en tableau indique si les données ont tendance à être manquantes pour plusieurs variables dans les observations individuelles. En d'autres termes, il vous permet de déterminer si les données sont manquantes conjointement.

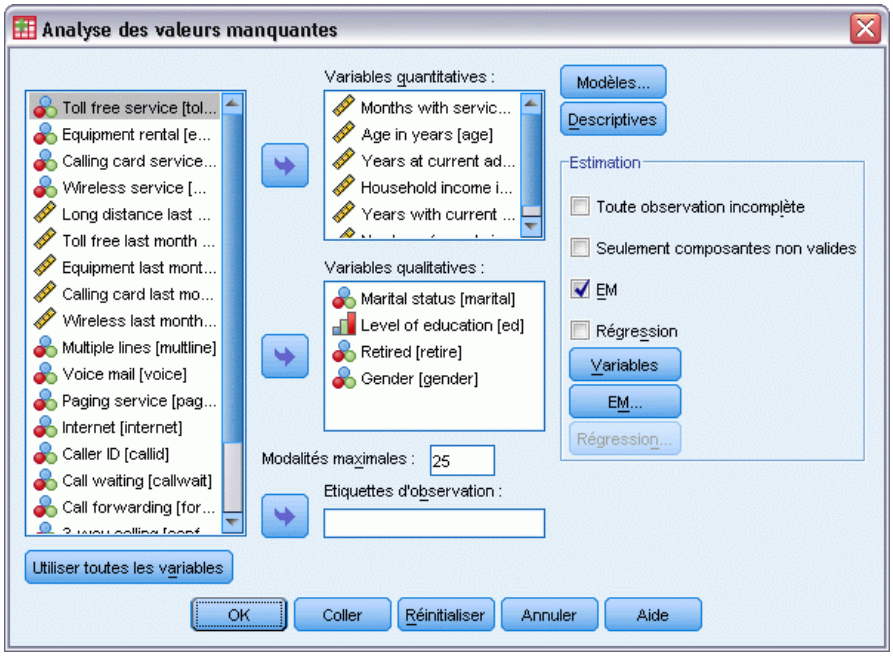
Trois modèles de données conjointement manquantes se produisent dans plus de 1 % des observations. Les variables*Nb d'années avec l'employeur actuel (employ)* et *retraite (retire)*sont conjointement manquantes plus souvent que les autres paires. Ce n'est pas surprenant, car *retraite* et *emploi* enregistrent des informations similaires. Si vous ignorez qu'un répondant est à la retraite, il est probable que vous ignoriez également depuis combien d'années le répondant travaille pour son employeur actuel.

Le *revenu (Revenu du ménage en milliers)* moyen semble varier considérablement en fonction du modèle de valeurs manquantes. En particulier, le *Revenu* moyen est beaucoup plus élevé pour 6 % (60 sur 1 000) des observations lorsque *marital (Situation familiale)* est manquant. (Il est également plus élevé lorsque *tenure (Nb de mois de service)* est manquant, mais ce modèle ne représente que 1,7 % des observations). Souvenez-vous que les personnes qui ont un niveau d'éducation plus élevé étaient moins disposées à répondre à la question portant sur la situation familiale. Cette tendance apparaît dans les effectifs affichés pour *ed (Niveau d'éducation).* Nous pourrions expliquer l'augmentation de la variable *revenu* en supposant que les personnes qui ont un niveau d'éducation plus élevé gagnent plus d'argent et sont moins susceptibles d'indiquer leur situation familiale.

Les statistiques descriptives et les modèles de données manquantes nous amènent à conclure que les données ne sont pas des valeurs manquantes complètement aléatoires. Nous pouvons confirmer cette conclusion à l'aide du test MCAR Little, affiché conjointement avec les estimations EM.

Réexécution de l'analyse pour le test MCAR Little

Figure 4-12
Boîte de dialogue Analyse des valeurs manquantes



- Rappeler la boîte de dialogue Descriptives de l'analyse des valeurs manquantes.
- Cliquez sur EM.
- Cliquez sur OK.

Figure 4-13
Tableau Moyennes EM

MonthsWithService	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
36,12	41,91	11,58	77,3941	11,22	2,29

a. Test MCAR : Khi-deux = 179,836, DDL = 107, Sig. = ,000

Les résultats du test MCAR Little apparaissent dans des notes de bas de page ajoutées à chaque tableau d'estimations EM. L'hypothèse nulle du test MCAR est que les données sont des valeurs manquantes complètement aléatoires (MCAR). Les données sont de type MCAR lorsque le patron des valeurs manquantes ne dépend pas des valeurs de données. Etant donné que la valeur de signification est inférieure à 0,05 dans notre exemple, nous pouvons conclure que les données *ne sont pas* des valeurs manquantes complètement aléatoires. Cela confirme la conclusion tirée des statistiques descriptives et des modèles mis en tableau.

À ce stade, comme les données ne sont pas des valeurs manquantes complètement aléatoires, il n'est pas recommandé de lister les observations contenant des valeurs manquantes, ou d'imputer séparément les valeurs manquantes. Cependant, vous pouvez utiliser l'imputation multiple pour analyser en détail cet ensemble de données.

Chapitre
5

Imputation multiple

Utilisation de l'imputation multiple pour compléter et analyser un ensemble de données

Un fournisseur de services de télécommunication souhaite mieux comprendre les types d'utilisation des services dans sa base de données client. Il dispose de données complètes sur les services utilisés par les clients, mais les informations démographiques collectées par l'entreprise comportent certaines valeurs manquantes. De plus, ces valeurs ne sont pas manquantes de façon complètement aléatoire. Par conséquent, l'imputation multiple sera utilisée pour compléter l'ensemble de données.

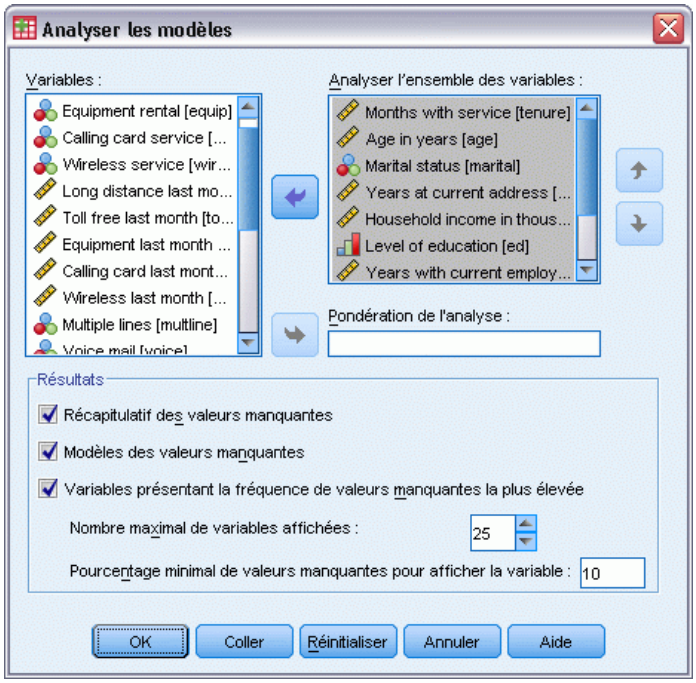
Un échantillon aléatoire issu de la base de données client figure dans le fichier *telco_missing.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Missing Values 19*.

Analyse des modèles de valeurs manquantes

▶ Tout d'abord, examinez les différents modèles des valeurs manquantes. A partir des menus, sélectionnez :

Analyse > Imputation multiple > Analyser les modèles...

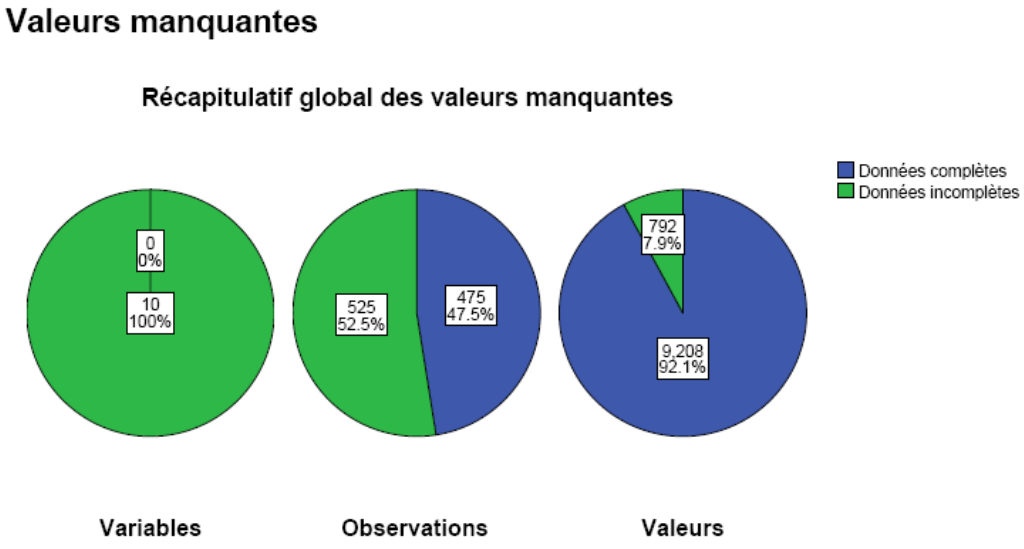
Figure 5-1
Boîte de dialogue Analyser les modèles



- Sélectionnez *Mois de service [tenure]* et *Nb de personnes dans le ménage [reside]* comme variables d'analyse.

Récapitulatif général

Figure 5-2
Récapitulatif général des valeurs manquantes



Le récapitulatif général des valeurs manquantes affiche trois diagrammes en secteurs qui présentent des aspects différents des valeurs manquantes dans les données.

- Le diagramme *Variables* indique que chacune des 10 variables d'analyse contient au moins une valeur manquante pour une observation.
- Le diagramme *Observations* indique que 525 des 1000 observations contiennent au moins une valeur manquante pour une variable.
- Le diagramme *Valeurs* indique que 792 des 10 000 valeurs (observations × variables) sont manquantes.

En moyenne, chaque observation contenant des valeurs manquantes contient des valeurs manquantes sur environ 1,5 variable sur 10. Ceci indique que l' **élimination des observations incomplètes** supprimerait de nombreuses informations dans l'ensemble de données.

Récapitulatif de variables

Figure 5-3
Récapitulatif de variables

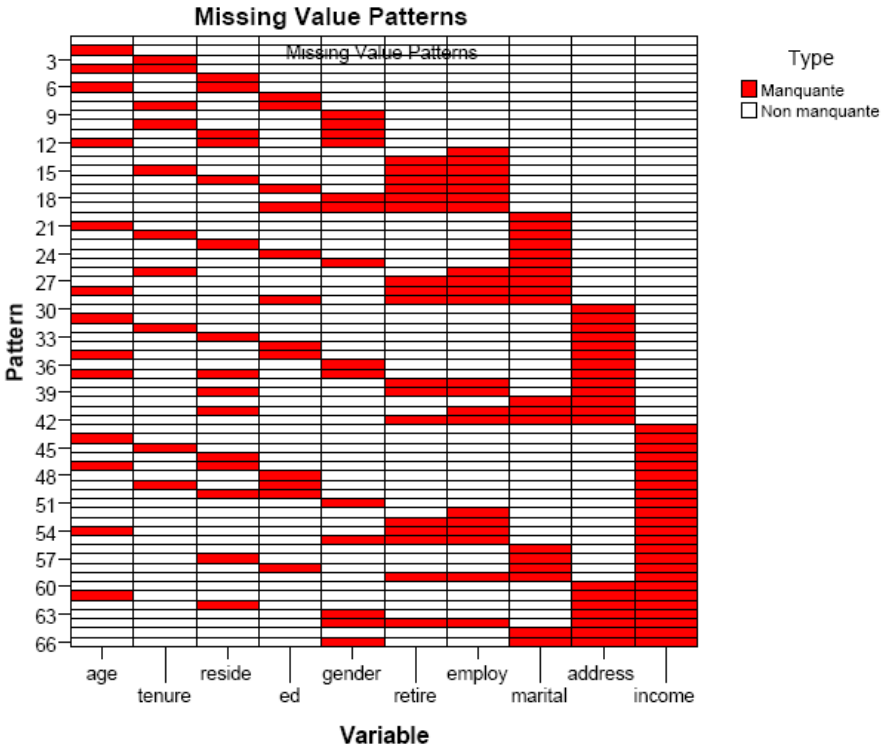
Récapitulatif de variables[a,b]

	Manquante				
	N	Pourcentage	N valide	Moyenne	Erreur Ecart
Household income in …	179	17,9%	821	71,1462	83,14424
Years at current address	150	15,0%	850	11,47	9,965
Marital status	115	11,5%	885		

Le récapitulatif de variables apparaît pour les variables contenant au moins 10% de valeurs manquantes et indique le nombre et le pourcentage de valeurs manquantes pour chaque variable du tableau. Il affiche également la moyenne et l'écart-type des valeurs valides des variables d'échelle et le nombre de valeurs valides pour toutes les variables. *Revenu du ménage en milliers, Nb d'années à la même adresse* et *Situation familiale* sont les variables contenant le plus de valeurs manquantes, dans cet ordre.

Modèles

Figure 5-4
Modèles de valeurs manquantes

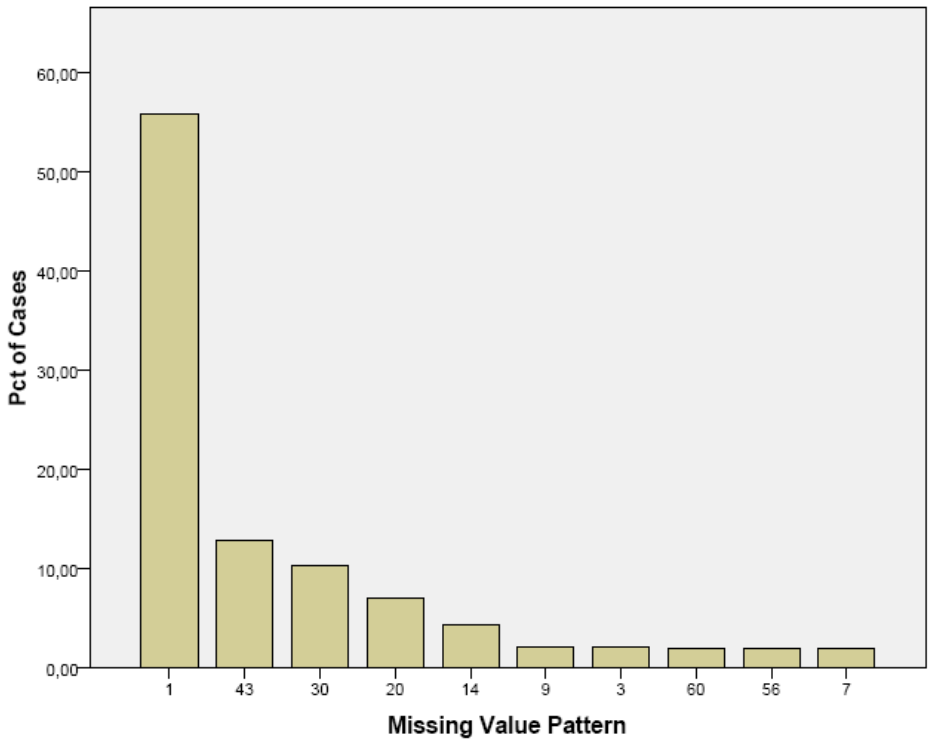


Le diagramme des modèles affiche les modèles des valeurs manquantes pour les variables d'analyse. Chaque modèle correspond à un groupe d'observations avec le même modèle de données complètes et incomplètes. Par exemple, le Modèle 1 représente des observations ne contenant aucune valeur manquante, alors que le Modèle 33 représente des observations contenant des valeurs manquantes sur *reside (Nb de personnes dans le ménage)* et *address (Nb d'années à la même adresse)*, et le Modèle 66 représente des observations contenant des valeurs manquantes sur *gender (Sexe)*, *marital (Situation familiale)*, *address* et *income (Revenu du ménage en milliers)*. Il est possible qu'un ensemble de données contiennent 2 modèles de $^{\text{nombre de variables}}$. Pour 10 variables d'analyse, cela donne $2^{10}=1024$; mais seuls 66 modèles sont représentés dans les 1000 observations de l'ensemble de données.

Le diagramme trie les variables d'analyse et les modèles de manière à révéler la monotonie, lorsqu'elle existe. Plus précisément, les variables sont triées de gauche à droite par ordre croissant des valeurs manquantes. Les modèles sont ensuite classés d'abord en fonction de la dernière variable (valeurs non manquantes puis valeurs manquantes), puis en fonction de la deuxième à la dernière variable, et ainsi de suite, de droite à gauche. Cela permet de déterminer si la méthode d'imputation monotone peut être utilisée pour vos données, ou dans le cas contraire, si vos données sont proches d'un modèle monotone. Si les données sont monotones, alors toutes les cellules manquantes et non manquantes du diagramme seront contiguës, c'est-à-dire qu'il n'y aura pas d'"îlots" de cellules non manquantes dans la partie inférieure droite du diagramme ni d'"îlots" de cellules manquantes dans la partie supérieure gauche du diagramme.

Cet ensemble de données n'est pas monotone et de nombreuses valeurs devraient être imputées afin d'obtenir la monotonie.

Figure 5-5
Effectifs des modèles



Lorsque des modèles sont demandés, un diagramme en bâtons affiche le pourcentage d'observations pour chaque modèle. Cela indique que plus de la moitié des observations dans l'ensemble de données suit le Modèle 1 et le diagramme des modèles de valeurs manquantes indique qu'il s'agit du modèle pour les observations sans valeurs manquantes. Le Modèle 43 représente les observations avec valeur manquante sur *income*, le Modèle 30 représente les observations avec valeur manquante sur *address* et le Modèle 20 représente les observations avec valeur manquante sur *marital*. La grande majorité des observations, environ 4 sur 5, est représentée par ces quatre modèles. Les modèles 14, 60 et 56 sont les seuls modèles parmi les dix modèles les plus fréquents à représenter des observations avec valeurs manquantes sur plus d'une variable.

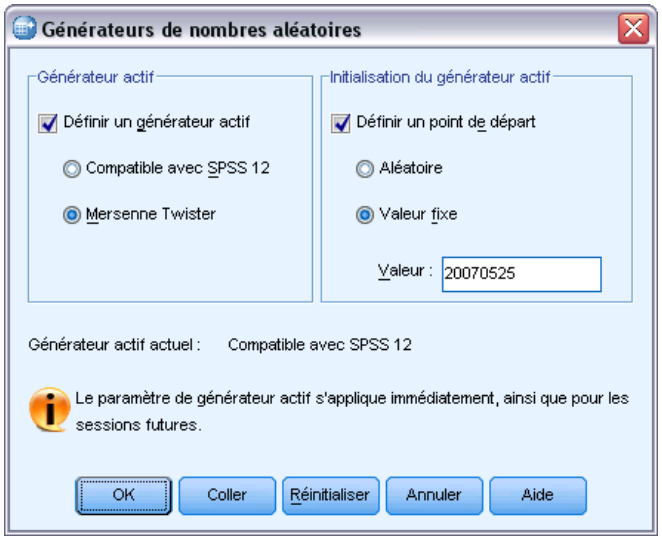
L'analyse des modèles manquants n'a pas révélé d'obstacles particuliers à l'imputation multiple, si ce n'est que la méthode monotone ne pourra pas vraiment être utilisée.

Imputation automatique des valeurs manquantes

Vous êtes maintenant prêt à imputer des valeurs ; nous commencerons par une exécution avec les paramètres automatiques mais avant de demander les imputations, nous définirons le générateur aléatoire. Définir le générateur aléatoire vous permet de reproduire l'analyse exactement.

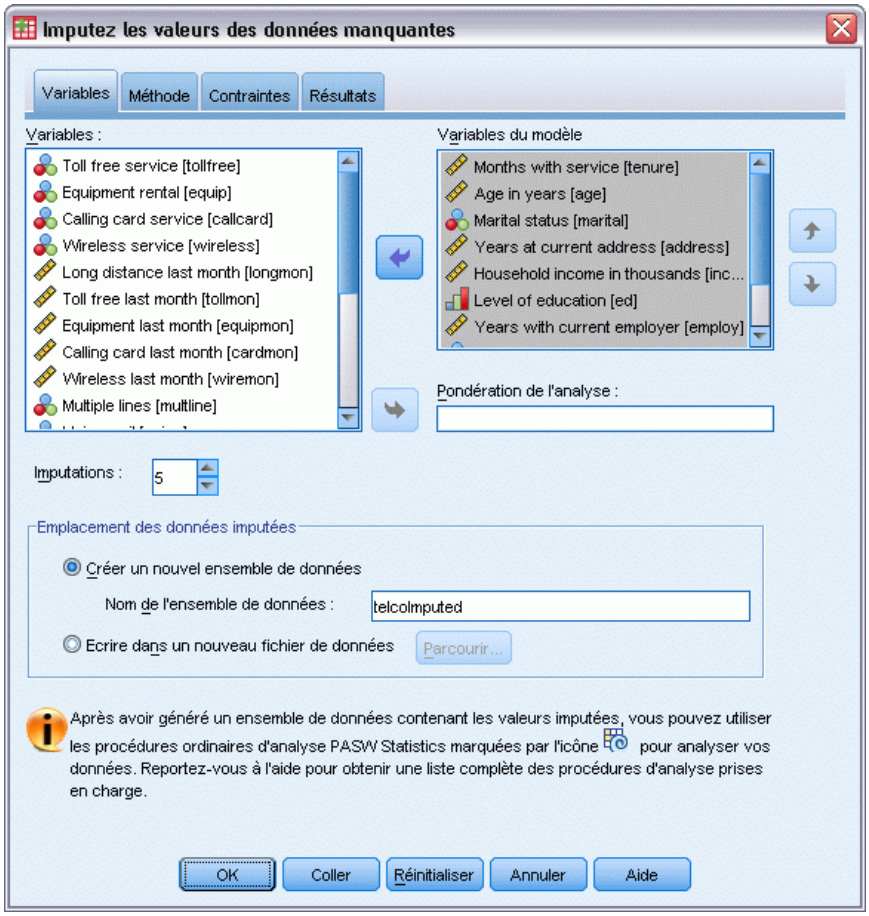
- Pour définir le générateur aléatoire, à partir des menus, sélectionnez :
  Transformer > Générateurs de nombres aléatoires…

Figure 5-6
Boîte de dialogue Générateurs de nombres aléatoires



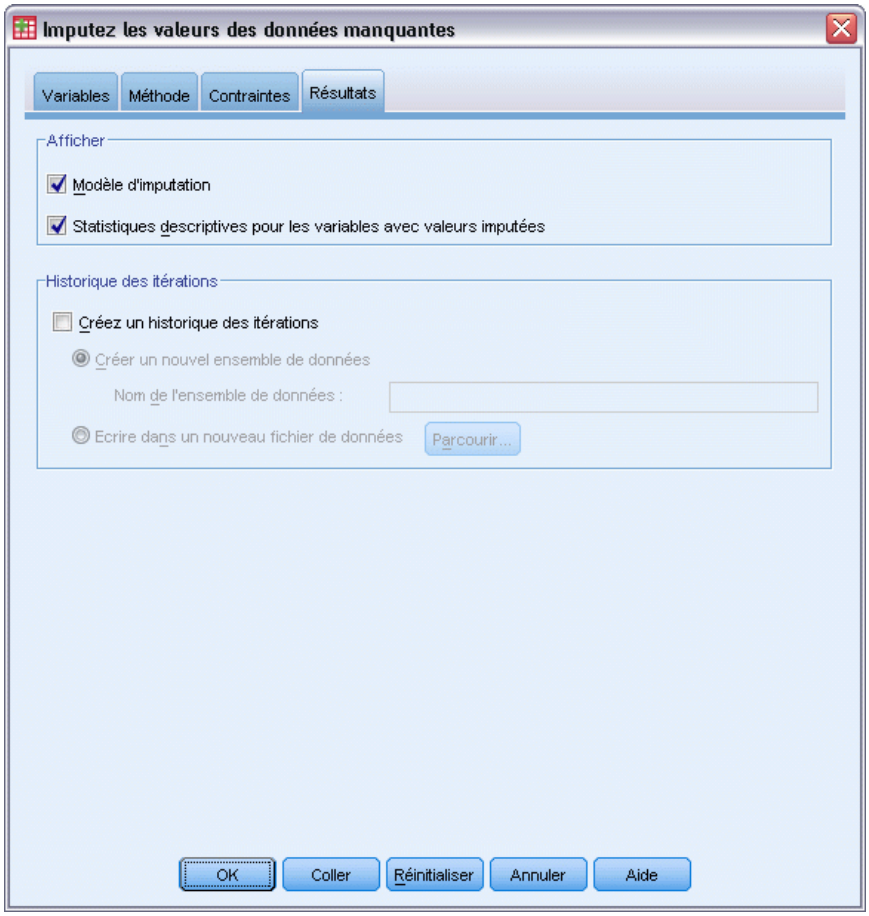
- Sélectionnez Définir un générateur actif.
- Sélectionnez Mersenne Twister.
- Sélectionnez Définir un point de départ.
- Sélectionnez Valeur fixe et tapez la valeur 20070525.
- Cliquez sur OK.
- Pour imputer les valeurs de données manquantes, dans les menus sélectionnez :
  Analyse > Imputation multiple > Imputer les valeurs de données manquantes...

Figure 5-7
Boîte de dialogue Imputer les valeurs de données manquantes



- Sélectionnez *Nb de mois de service [tenure]* et *Nb de personnes dans le ménage [reside]* comme variables du modèle d'imputation.
- Saisissez telcoImputed comme ensemble de données d'enregistrement des données imputées.
- Cliquez sur l'onglet Résultats.

Figure 5-8
Onglet Résultats



- Sélectionnez Statistiques descriptives pour variables avec valeurs imputées.
- Cliquez sur OK.

Spécifications des imputations

Figure 5-9
Spécifications des imputations

Méthode d'imputation	Automatique
Nombre d'imputations	5
Modèle pour variables ...	Régression linéaire
Intéractions inclues da...	(aucun)
Pourcentage maximal ...	100,0%

Le tableau des spécifications des imputations est une présentation utile des demandes effectuées permettant de confirmer que les spécifications étaient correctes.

Résultats des imputations

Figure 5-10
Résultats des imputations

Méthode d'imputation		Spécification entièrement conditionnelle
Itérations de méthode de spécification entièrement conditionnelle		10
Variables dépendantes	Imputée	tenure,age,marital,address,income,ed, employ,retire,gender,reside
	Non imputée (valeurs manquantes trop nombreuses)	
	Non imputée (aucune valeur manquante)	
Séquence d'imputation		age,tenure,reside,ed,gender,retire,employ, marital,address,income

Les résultats des imputations sont une présentation de ce qui s'est passé pendant le processus d'imputation. Veuillez noter les points suivants :

- La méthode d'imputation dans le tableau des spécifications était Automatique et la méthode choisie par la sélection de méthodes automatique était Spécification entièrement conditionnelle.
- Toutes les variables demandées ont été imputées.
- La séquence d'imputation est dans le même ordre que celui dans lequel les variables apparaissent sur l'axe *x* dans le diagramme des modèles de valeurs manquantes.

Modèles d'imputation

Figure 5-11
modèles d'imputation

	Modèle		Valeurs manquantes	Valeurs imputées
	Type	Effets		
Age in years	Régression linéaire	ed,gender,retire,marital,tenure,reside, employ,address,income	25	125
Months with service	Régression linéaire	ed,gender,retire,marital,age,reside, employ,address,income	32	160
Number of people in household	Régression linéaire	ed,gender,retire,marital,age,tenure, employ,address,income	34	170
Level of education	Régression logistique	gender,retire,marital,age,tenure,reside, employ,address,income	35	175
Gender	Régression logistique	ed,retire,marital,age,tenure,reside, employ,address,income	42	210
Retired	Régression logistique	ed,gender,marital,age,tenure,reside, employ,address,income	84	420
Years with current employer	Régression linéaire	ed,gender,retire,marital,age,tenure, reside,address,income	96	480
Marital status	Régression logistique	ed,gender,retire,age,tenure,reside, employ,address,income	115	575
Years at current address	Régression linéaire	ed,gender,retire,marital,age,tenure, reside,employ,income	150	750
Household income in thousands	Régression linéaire	ed,gender,retire,marital,age,tenure, reside,employ,address	179	895

Le tableau des modèles d'imputation présente des détails supplémentaires sur la façon dont chaque variable a été imputée. Veuillez noter les points suivants :

- Les variables apparaissent dans l'ordre de la séquence d'imputation.
- Les variables d'échelle sont modélisées avec une régression linéaire et les variables catégorielles avec une régression logistique.
- Chaque modèle utilise toutes les autres variables comme effets principaux.
- Le nombre de valeurs manquantes pour chaque variable est répertorié, avec le nombre total de valeurs imputées pour cette variable (nombre de valeurs manquantes × nombre d'imputations).

Statistiques descriptives

Figure 5-12
Statistiques descriptives pour la variable tenure (Nb de mois de service)

Données	Im...	N	Moyenne	Erreur Ecart	Minimum	Maximum
Données initiales		968	35,56	21,268	1,00	72,00
Valeurs imputées	1	32	39,17	21,610	1,52	93,69
	2	32	36,63	16,958	4,29	85,29
	3	32	42,39	23,733	4,82	90,89
	4	32	39,97	21,813	10,43	95,04
	5	32	41,86	21,703	5,18	88,84
Données complètes après imputation	1	1000	35,68	21,278	1,00	93,69
	2	1000	35,60	21,138	1,00	85,29
	3	1000	35,78	21,372	1,00	90,89
	4	1000	35,70	21,289	1,00	95,04
	5	1000	35,76	21,300	1,00	88,84

Les tableaux des statistiques descriptives présentent des récapitulatifs des variables avec valeurs imputées. Un tableau séparé est produit pour chaque variable. Les types de statistiques affichés dépendent du type de la variable (d'échelle ou catégorielle).

Les statistiques pour les variables d'échelle comprennent l'effectif, la moyenne, l'écart-type, le minimum et le maximum, pour les données d'origine, chaque ensemble de valeurs imputées et chaque ensemble de données complet (conjuguant les données d'origine et les valeurs imputées).

Le tableau des statistiques descriptives pour *tenure (Nb de mois de service)* présente les moyennes et les écarts-types dans chaque ensemble de valeurs imputées, quasiment égaux à ceux des données d'origine ; mais un problème surgit lorsqu'on examine le minimum et que des valeurs négatives pour *tenure* ont été imputées.

Figure 5-13
Statistiques descriptives pour la variable marital (Situation familiale)

Données	Im...	Mo...	N	Pourcentage
Données initiales		0	456	51,5
		1	429	48,5
Valeurs imputées	1	0	51	44,3
		1	64	55,7
	2	0	49	42,6
		1	66	57,4
	3	0	51	44,3
		1	64	55,7
	4	0	50	43,5
		1	65	56,5
	5	0	54	47,0
		1	61	53,0
Données complètes après imputation	1	0	507	50,7
		1	493	49,3
	2	0	505	50,5
		1	495	49,5
	3	0	507	50,7
		1	493	49,3
	4	0	506	50,6
		1	494	49,4
	5	0	510	51,0
		1	490	49,0

Pour les variables catégorielles, les statistiques comprennent l'effectif et le pourcentage par catégorie pour les données d'origine, les valeurs imputées et les données complètes. Le tableau pour *marital (Situation familiale)* contient un résultat intéressant car pour les valeurs imputées, la proportion des observations évaluées comme mariées est plus importante que celle des données d'origine. Ceci peut provenir d'une variation aléatoire, ou le risque de valeur manquante peut être lié à la valeur de cette variable.

Figure 5-14
Statistiques descriptives pour la variable income (Revenu du ménage en milliers)

Données	Im...	N	Moyenne	Erreur Ecart	Minimum	Maximum
Données initiales		821	71,1462	83,14424	9,0000	944,0000
Valeurs imputées	1	179	116,4104	74,76155	1,1206	283,4055
	2	179	115,2794	81,58019	1,6164	372,4810
	3	179	119,4888	77,46842	1,0957	392,0048
	4	179	111,8589	74,05791	1,0153	335,2312
	5	179	105,6987	77,29500	,2568	434,6000
Données complètes après imputation	1	1000	79,2485	83,49607	1,1206	944,0000
	2	1000	79,0460	84,53795	1,6164	944,0000
	3	1000	79,7995	84,18673	1,0957	944,0000
	4	1000	78,4337	83,03836	1,0153	944,0000
	5	1000	77,3311	83,15323	,2568	944,0000

Comme *tenure*, et toutes les autres variables d'échelle, *income (Revenu du ménage en milliers)* présente des valeurs imputées négatives — nous aurons donc besoin d'exécuter un modèle personnalisé avec des contraintes sur certaines variables. Cependant, *income* présente d'autres problèmes potentiels. Les valeurs moyennes pour chaque imputation sont considérablement plus

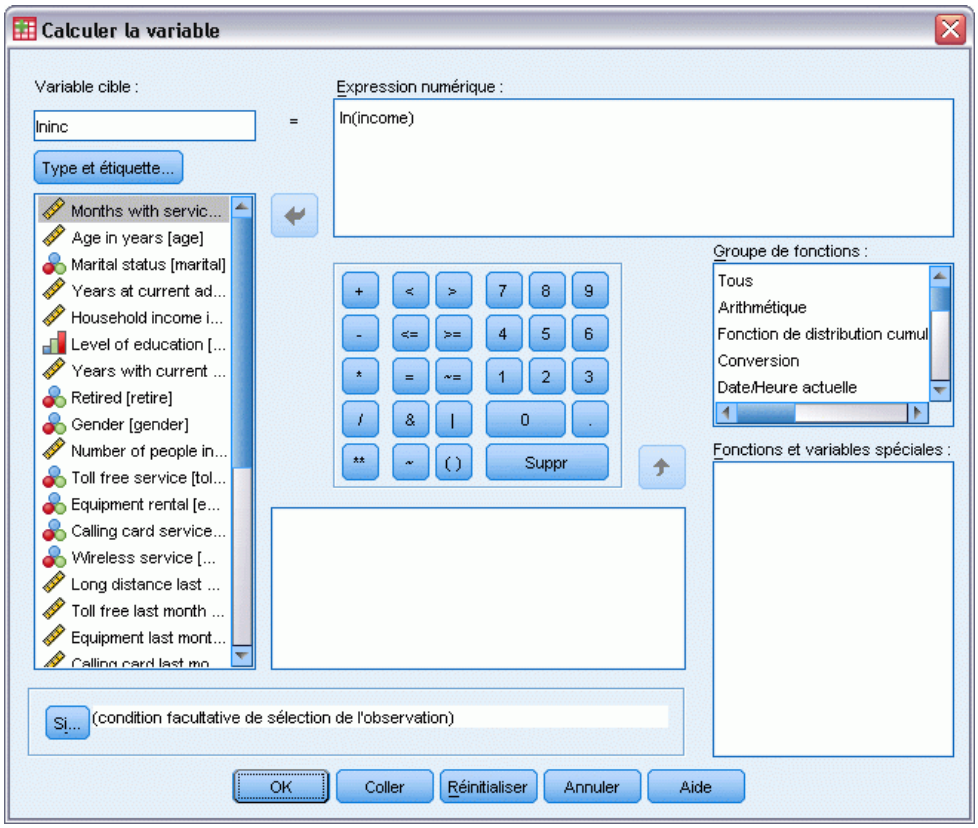
élevées que pour les données d'origine et les valeurs maximales pour chaque imputation sont considérablement moins élevées que pour les données d'origine. La répartition du revenu a tendance à être fortement asymétrique, ce qui pourrait être la cause du problème.

Modèle d'imputation personnalisé

Afin d'éviter que les valeurs imputées ne sortent de la plage de valeurs raisonnable pour chaque variable, nous spécifierons un modèle d'imputation personnalisé avec des contraintes sur les variables. De plus, *Revenu du ménage en milliers* est fortement asymétrique et des analyses supplémentaires utiliseront probablement le logarithme du *revenu*. Il paraît donc cohérent d'imputer directement le log du revenu.

- Vérifiez que l'ensemble de données d'origine est actif.
- Pour créer une variable log du revenu, à partir des menus, sélectionnez :
  Transformer > Calculer la variable...

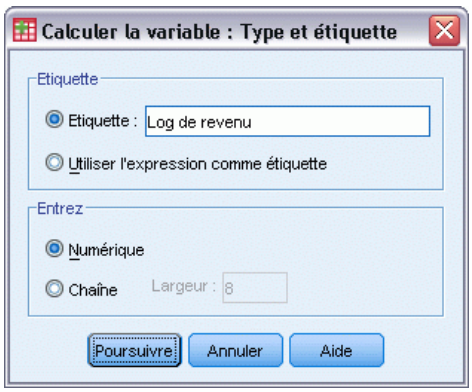
Figure 5-15
Boîte de dialogue Calculer la variable



- Tapez *lninc* comme variable de destination.
- Entrez l'expression numérique ln(Income).

▶ Cliquer sur Type & Etiquette..

Figure 5-16
Boîte de dialogue Type et étiquette

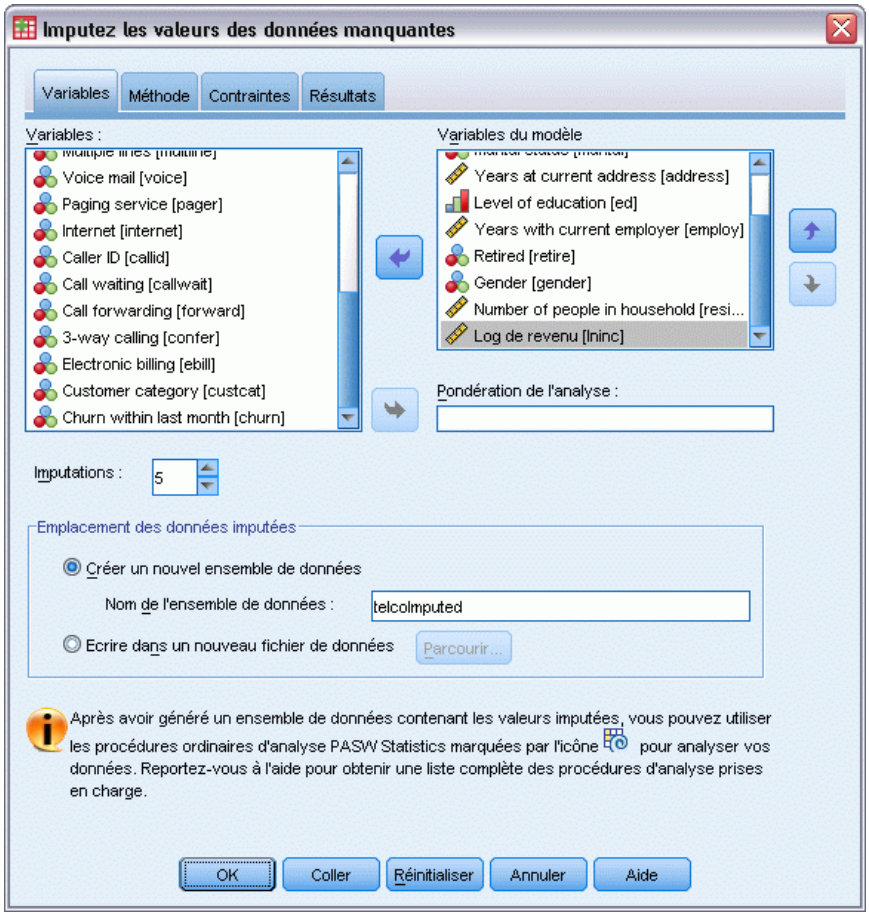


▶ Saisissez *Log de revenu* comme étiquette.

▶ Cliquez sur Poursuivre.

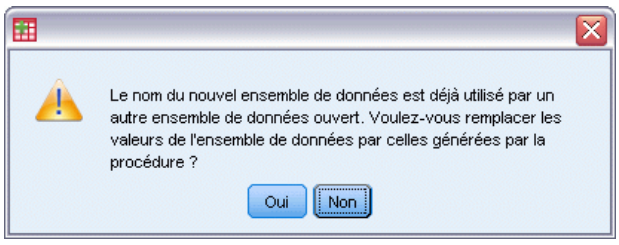
▶ Cliquez sur OK dans la boîte de dialogue Calculer la variable.

Figure 5-17
Onglet Variables avec log de revenu remplaçant Revenu du ménage en milliers dans le modèle d'imputation



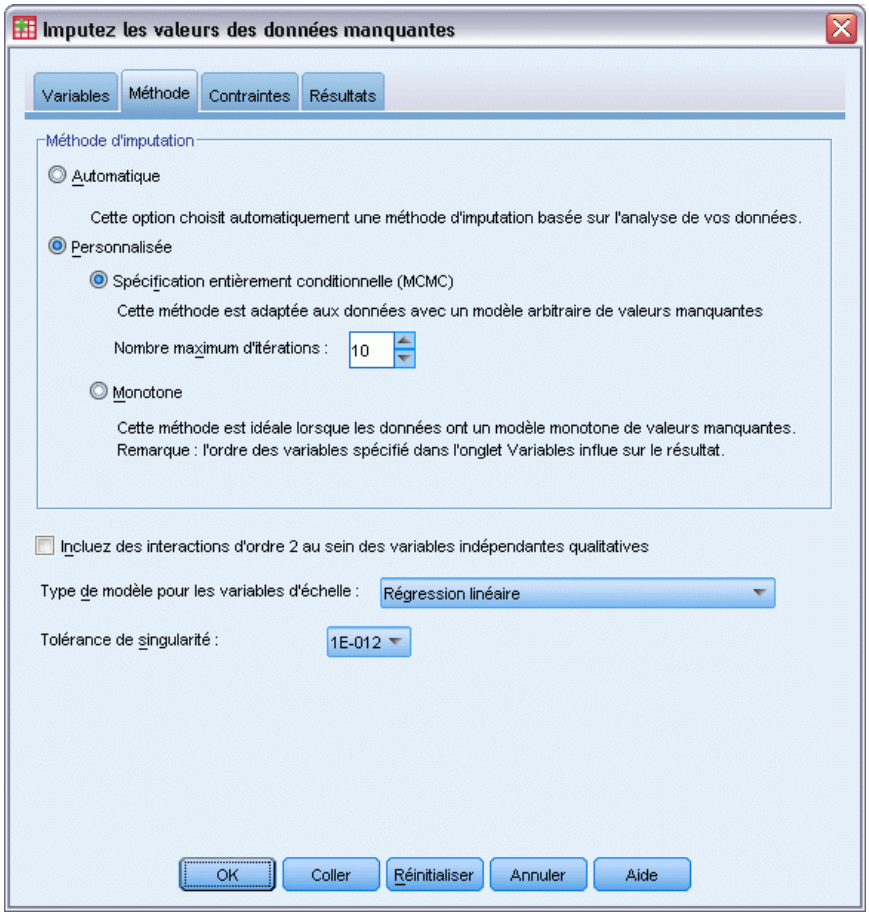
- Rappelez la boîte de dialogue Imputer les valeurs de données manquantes et cliquez sur l'onglet Variables.
- Désélectionnez *Revenu du ménage en milliers [income]* et sélectionnez *Log de revenu [lninc]* comme variables dans le modèle.
- Cliquez sur l'onglet Méthode.

Figure 5-18
Alerte de remplacement d'un ensemble de données existant



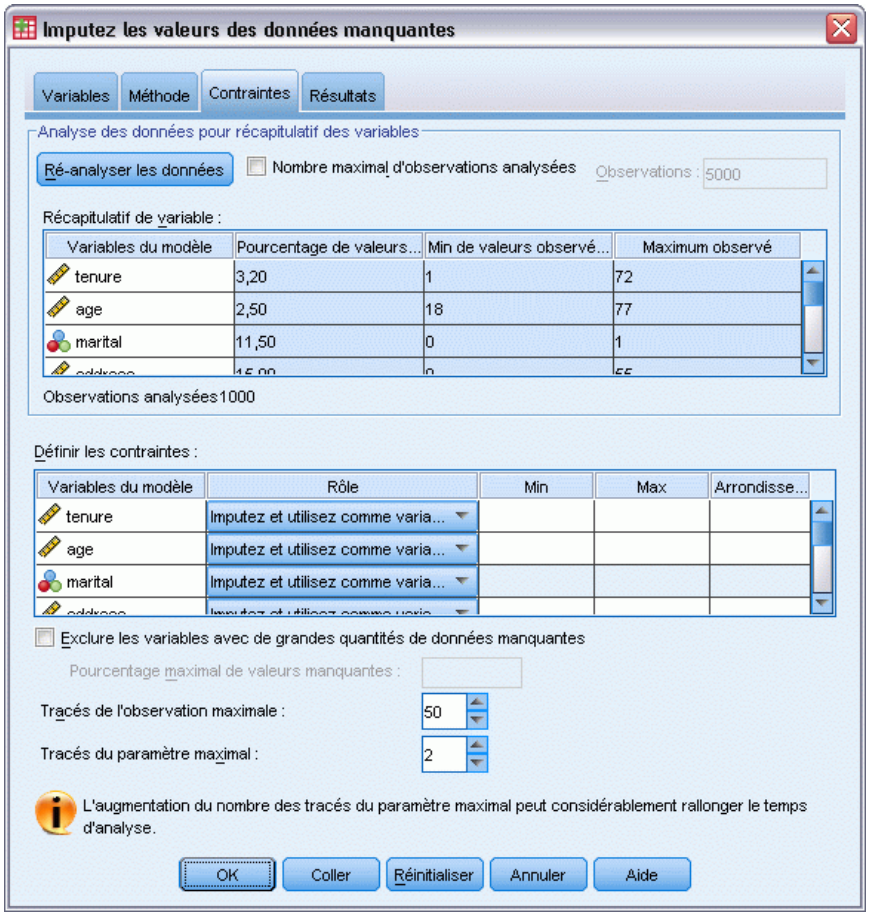
- Cliquez sur Oui dans l'alerte affichée.

Figure 5-19
Onglet Méthode



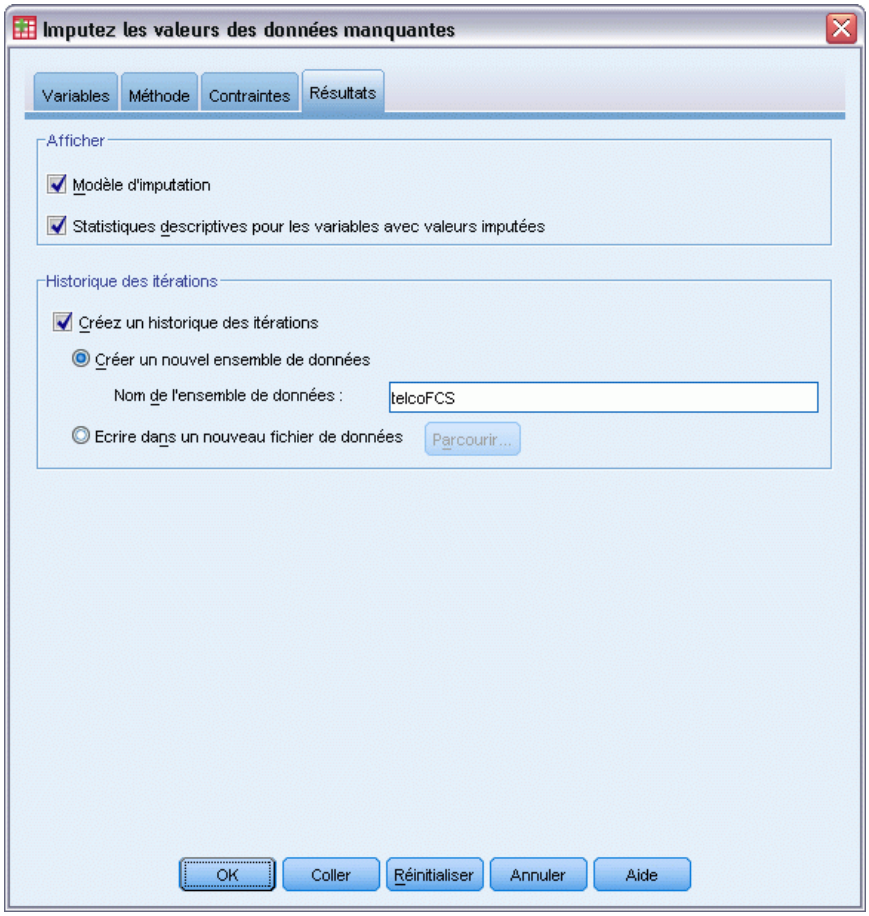
- Sélectionnez Personnalisé et laissez Spécification entièrement conditionnelle sélectionnée comme méthode d'imputation.
- Cliquez sur l'onglet Contraintes.

Figure 5-20
Onglet Contraintes



- Cliquez sur Analyser les données.
- Dans la grille Définir les contraintes, saisissez 1 comme valeur minimale pour *Nb de mois de service [tenure]*.
- Saisissez 18 comme valeur minimale pour *age (Age en années)*.
- Saisissez 0 comme valeur minimale pour *address (Nombre d'années à la même adresse)*.
- Saisissez 0 comme valeur minimale pour *employ (Nombre d'années avec l'employeur actuel)*.
- Saisissez 1 comme valeur minimale et 1 comme niveau d'arrondi pour *reside (Nombre de personnes dans le ménage)*. Veuillez noter que bien que de nombreuses autres variables d'échelle sont répertoriées sous forme de valeurs entières, il est normal de déclarer que quelqu'un a vécu pendant 13,8 années à la même adresse mais beaucoup moins que 2,2 personnes y vivent.
- Saisissez 0 comme valeur minimale pour *Ininc (Log de revenu)*.
- Cliquez sur l'onglet Résultats.

Figure 5-21
Onglet Résultats



- Sélectionnez Créer un historique des itérations et saisissez telcoFCS comme nom du nouvel ensemble de données.
- Cliquez sur OK.

Contraintes d'imputation

Figure 5-22
Contraintes d'imputation

	Rôle dans l'imputation		Valeurs imputées		
	Dépendant	Variable indépendante	Minimum	Maximum	Arrondissement
Months with service	Oui	Oui	1	(aucun)	
Age in years	Oui	Oui	18	(aucun)	
Marital status	Oui	Oui			
Years at current address	Oui	Oui	0	(aucun)	
Level of education	Oui	Oui			
Years with current …	Oui	Oui	0	(aucun)	
Retired	Oui	Oui			
Gender	Oui	Oui			
Number of people in …	Oui	Oui	1	(aucun)	Entier
Log of Income	Oui	Oui	0	(aucun)	

Le modèle d'imputation personnalisé génère un nouveau tableau qui présente les contraintes placées sur le modèle d'imputation. Tout semble en accord avec vos spécifications.

Statistiques descriptives

Figure 5-23
Statistiques descriptives pour la variable tenure (Nb de mois de service)

Données	Im…	N	Moyenne	Erreur Ecart	Minimum	Maximum
Données initiales		968	35,56	21,268	1,00	72,00
Valeurs imputées	1	32	41,46	22,594	5,72	90,18
	2	32	37,00	18,747	2,41	66,86
	3	32	41,23	22,819	6,89	92,52
	4	32	38,21	21,789	2,44	85,74
	5	32	39,51	21,693	2,66	83,45
Données complètes après imputation	1	1000	35,75	21,325	1,00	90,18
	2	1000	35,61	21,185	1,00	72,00
	3	1000	35,74	21,331	1,00	92,52
	4	1000	35,65	21,279	1,00	85,74
	5	1000	35,69	21,282	1,00	83,45

Le tableau des statistiques descriptives pour *tenure (Nb de mois en service)* d'après le modèle d'imputation personnalisé avec contraintes indique que le problème des valeurs imputées négatives pour *tenure* a été résolu.

Figure 5-24
Statistiques descriptives pour la variable marital (Situation familiale)

Données	Im...	Mo...	N	Pourcentage
Données initiales		0	456	51,5
		1	429	48,5
Valeurs imputées	1	0	51	44,3
		1	64	55,7
	2	0	50	43,5
		1	65	56,5
	3	0	51	44,3
		1	64	55,7
	4	0	49	42,6
		1	66	57,4
	5	0	47	40,9
		1	68	59,1
Données complètes après imputation	1	0	507	50,7
		1	493	49,3
	2	0	506	50,6
		1	494	49,4
	3	0	507	50,7
		1	493	49,3
	4	0	505	50,5
		1	495	49,5
	5	0	503	50,3
		1	497	49,7

Le tableau pour *marital (Situation familiale)* a maintenant une imputation (3) dont la distribution est plus en accord avec les données d'origine, mais la majorité présente encore une proportion d'observations estimées comme mariées plus importante que celle des données d'origine. Ceci pourrait provenir d'une variation aléatoire mais nécessite un examen supplémentaire des données pour déterminer si ces valeurs ne sont pas manquantes de manière aléatoire (MAR). Nous n'étudierons pas ce problème plus avant.

Figure 5-25
Statistiques descriptives pour la variable lninc (Log de revenu)

Données	Im...	N	Moyenne	Erreur Ecart	Minimum	Maximum
Données initiales		821	3,9291	,75305	2,1972	6,8501
Valeurs imputées	1	179	4,2706	,85293	2,1093	6,3817
	2	179	4,2912	,84755	2,2327	6,8152
	3	179	4,1914	,84340	2,2035	6,6232
	4	179	4,1058	,86682	1,6335	6,0839
	5	179	4,1981	,92481	1,6239	6,1638
Données complètes après imputation	1	1000	3,9902	,78247	2,1093	6,8501
	2	1000	3,9939	,78279	2,1972	6,8501
	3	1000	3,9760	,77610	2,1972	6,8501
	4	1000	3,9607	,77714	1,6335	6,8501
	5	1000	3,9772	,79279	1,6239	6,8501

Comme *tenure*, et toutes les autres variables d'échelle, *lninc (Log de revenu)* ne présente pas de valeurs imputées négatives. De plus, les valeurs moyennes des imputations sont plus proches de la moyenne des données d'origine que dans l'exécution de l'imputation automatique — dans l'échelle *income*, la moyenne pour les données d'origine pour *lninc* est d'environ $e^{3,9291}$=50,86,

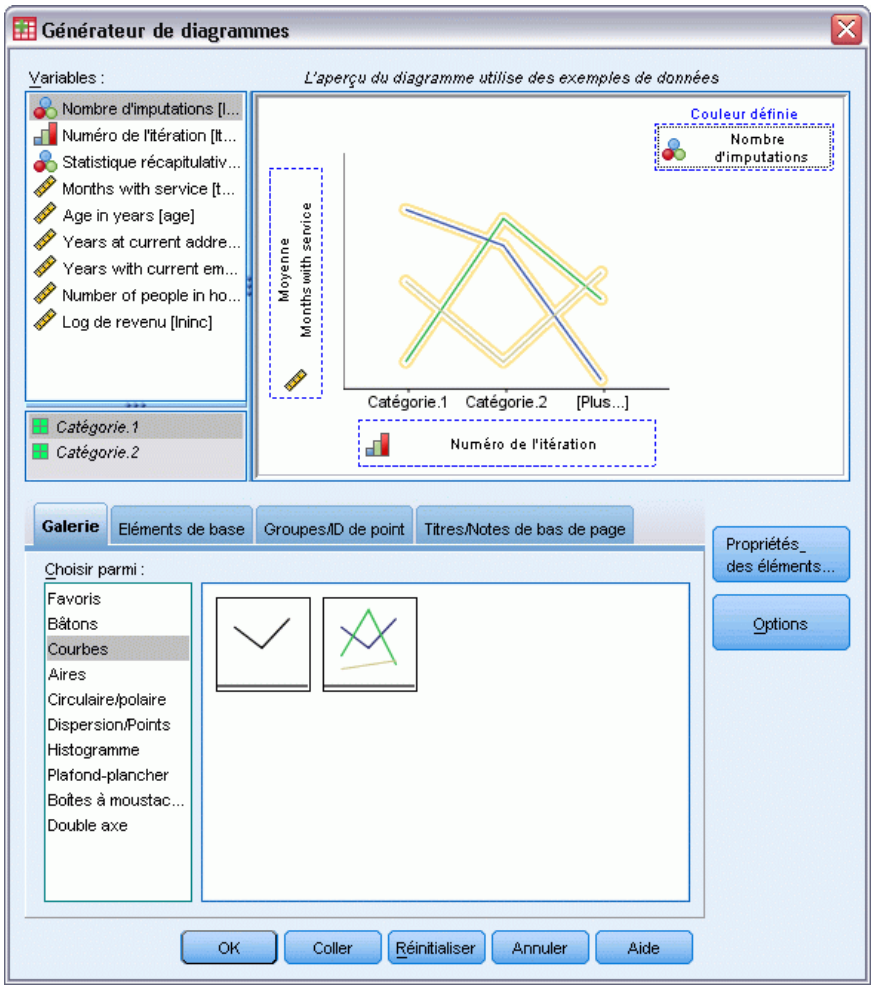
alors que la valeur moyenne habituelle dans les imputations est d'environ $e^{4,2}$=66,69. De plus, les valeurs maximales pour chaque imputation sont plus proches de la valeur maximale pour les données d'origine.

Vérification de la convergence FCS

Si la méthode de spécification entièrement conditionnelle est utilisée, il paraît sage de vérifier les diagrammes des moyennes et des écarts-types par itération et par imputation pour chaque variable d'échelle dépendante dont les valeurs sont imputées afin de mieux évaluer la convergence des modèles.

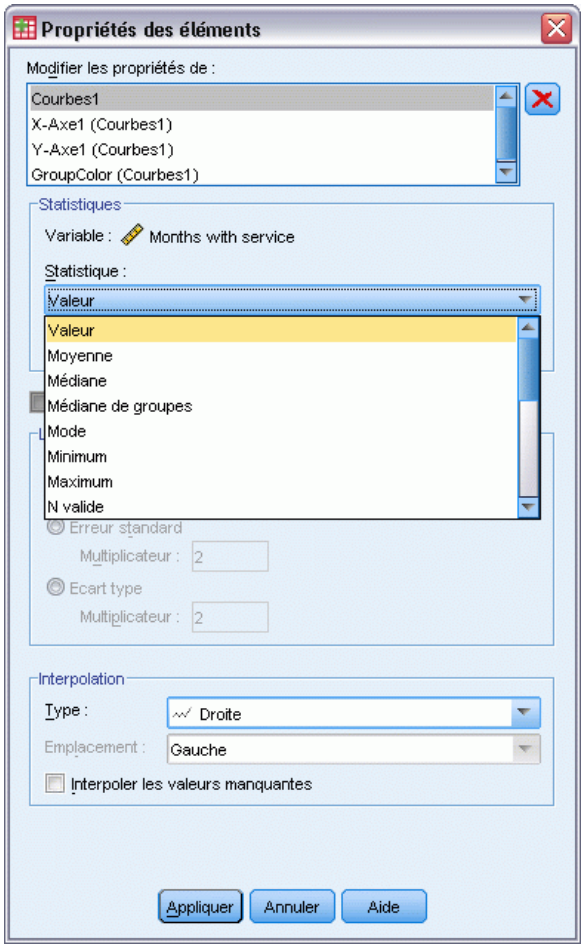
- Pour créer ce type de diagramme, activez l'ensemble de données *telcoFCS* puis parmi les menus, choisissez :
  Graphes > Générateur de diagrammes...

Figure 5-26
Générateur de diagrammes, diagramme à courbes multiples



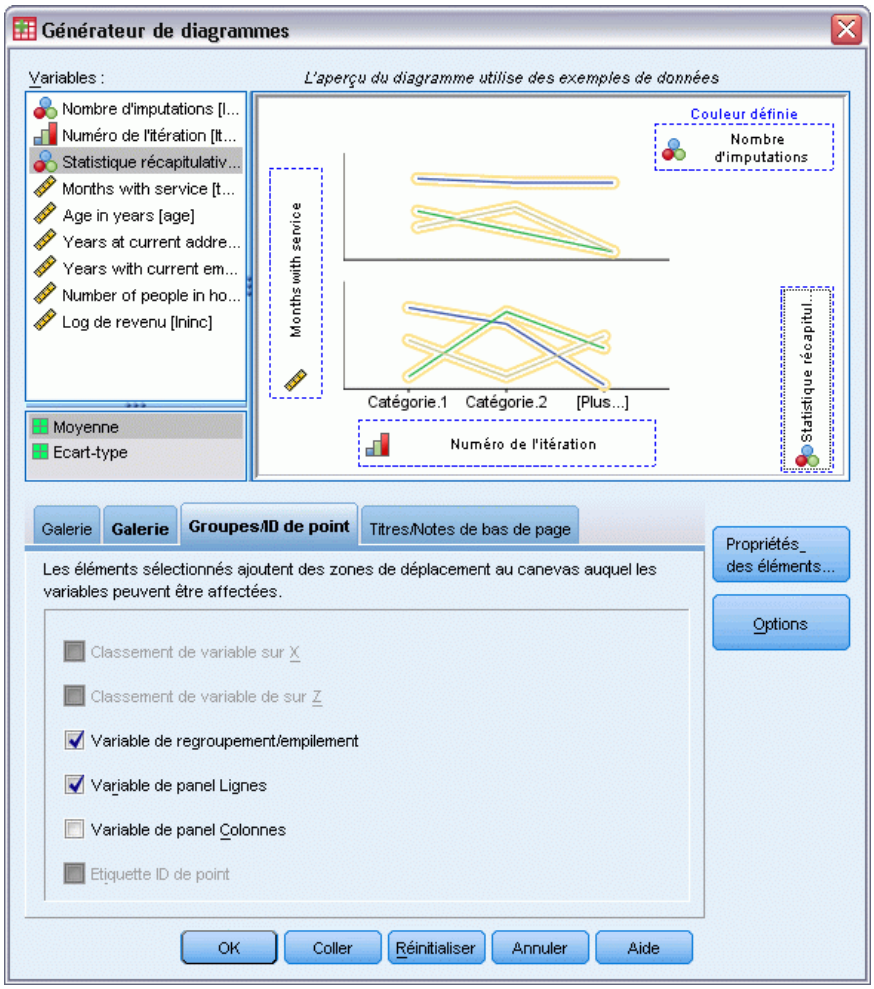
- Sélectionnez la galerie Courbe et choisissez Courbes multiples.
- Sélectionnez *Nb de mois avec service [tenure]* comme variable à tracer sur l'axe *Y*.
- Sélectionnez *Nombre d'itérations [Iteration_]* comme variable à tracer sur l'axe *X*.
- Sélectionnez *Nombre d'imputations [Imputations_]* comme variable d'après laquelle définir les couleurs.

Figure 5-27
Générateur de diagrammes, propriétés des éléments



- Dans les propriétés des éléments, sélectionnez Valeur comme statistique à afficher.
- Cliquez sur Appliquer.
- Dans le Générateur de diagrammes, cliquez sur l'onglet Groupes/ID de point.

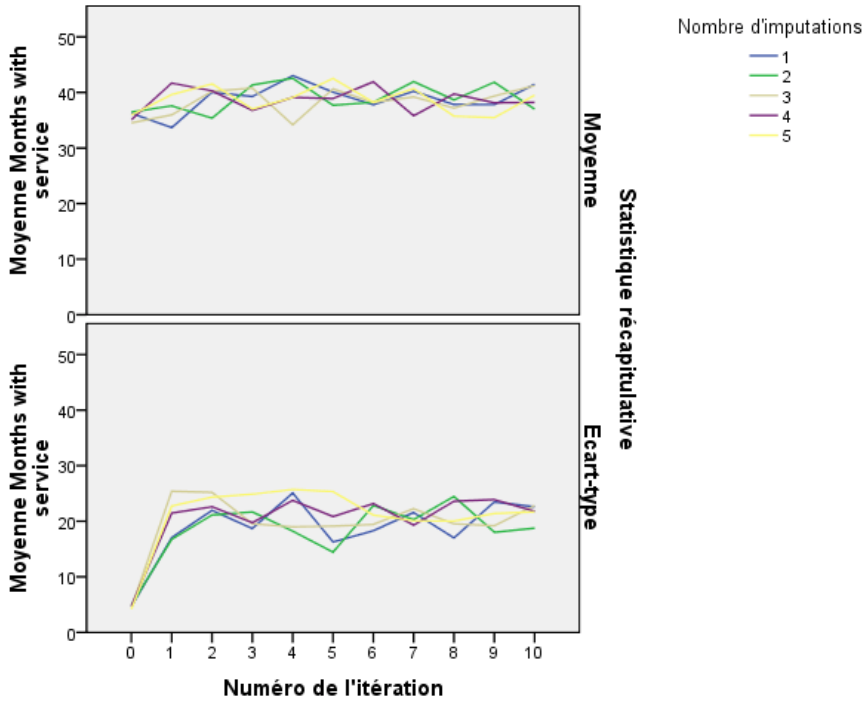
Figure 5-28
Générateur de diagrammes, onglet Groupes/ID de point



- Sélectionnez Variable de panel Lignes.
- Sélectionnez *Statistique récapitulative [SummaryStatistic_]* comme variable de panel.
- Cliquez sur OK.

Diagrammes de convergence FCS

Figure 5-29
Diagramme de convergence FCS



Vous avez créé une paire de diagrammes curvilignes multiples, indiquant la moyenne et l'écart-type des valeurs imputées de *Nb de mois avec service [tenure]* à chaque itération de la méthode d'imputation FCS pour chacune des 5 imputations appelées. L'objectif de ce diagramme est de rechercher des modèles dans les courbes. Il ne devrait y en avoir aucun et les courbes devraient être 'aléatoires'. Vous pouvez créer des diagrammes similaires pour les autres variables d'échelle. Vous noterez que ces diagrammes ne présentent aucun modèle perceptible.

Analyser les données complètes

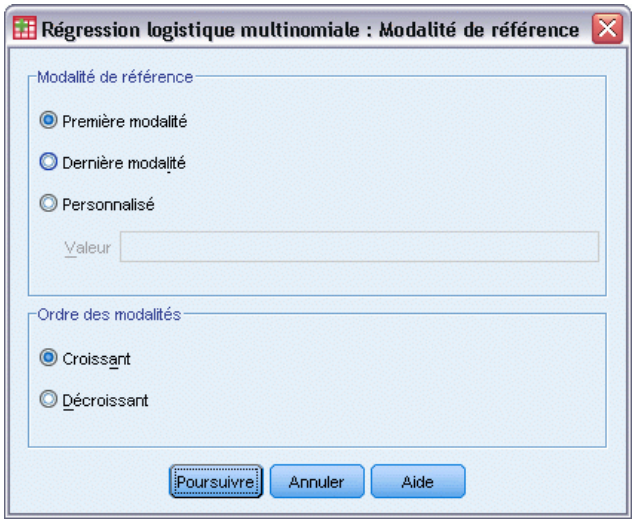
A présent que vos valeurs imputées semblent satisfaisantes, vous êtes prêt à exécuter une analyse sur les données "complètes". L'ensemble de données contient une variable *Catégorie de client [custcat]* qui segmente la base client par type d'utilisation des services en catégorisant les clients en quatre groupes. Si vous pouvez ajuster un modèle utilisant des informations démographiques pour prévoir les groupes d'affectation, vous pouvez personnaliser les offres pour chaque client éventuel.

- Activez l'ensemble de données *telcoImputed* . Pour créer un modèle de régression logistique multinomiale pour les données complètes, parmi les menus, sélectionnez :
  Analyse > Régression > Logistique multinomiale...

Figure 5-30
Boîte de dialogue Régression logistique multinomiale

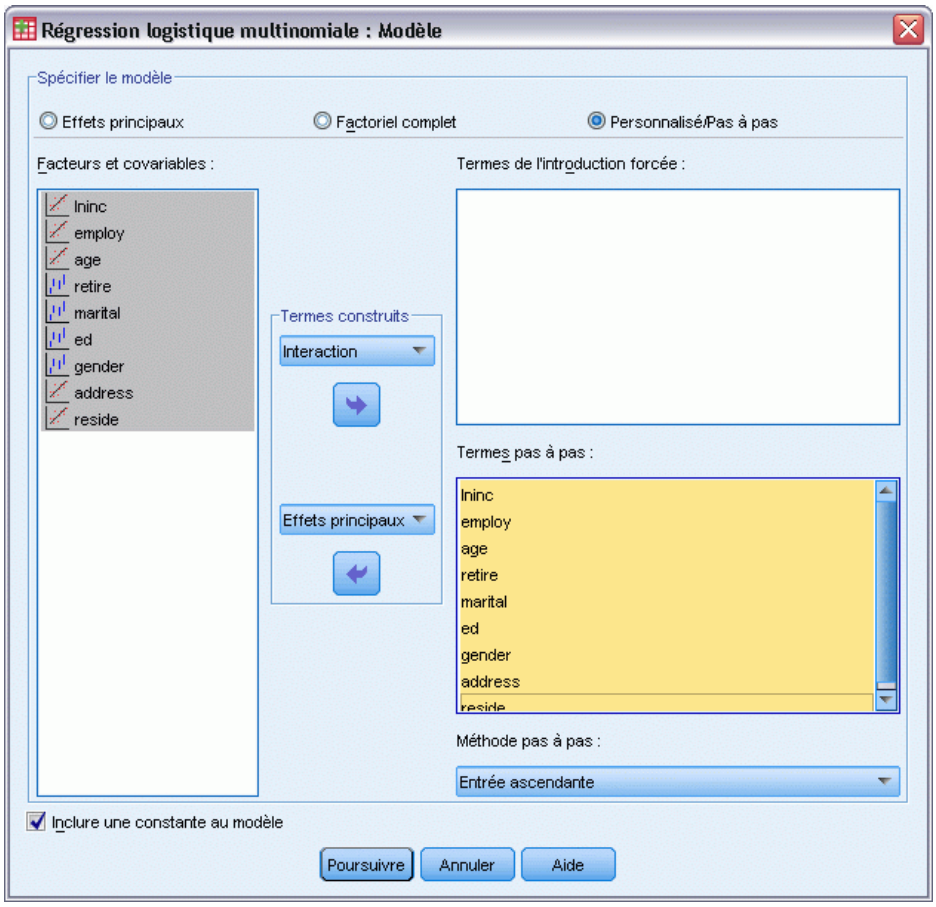
- Sélectionnez *Catégorie de client* comme variable dépendante.
- Sélectionnez *Situation familiale*, *Niveau d'éducation*, *Retraité* et *Sexe* comme facteurs.
- Sélectionnez *Age en années*, *Nb d'années à la même adresse*, *Nb d'années avec l'employeur actuel*, *Nombre de personnes dans le ménage* et *Log de revenu* comme covariables.
- Pour comparer les autres clients à ceux qui ont souscrit au service de base, sélectionnez donc *Catégorie de client* et cliquez sur Modalité de référence.

Figure 5-31
Boîte de dialogue Modalité de référence



- Sélectionnez Première modalité.
- Cliquez sur Poursuivre.
- Cliquez sur Modèle dans la boîte de dialogue Régression logistique multinomiale.

Figure 5-32
Boîte de dialogue Modèle



- Sélectionnez Personnalisé/Pas à pas.
- Sélectionnez Effets principaux dans la liste déroulante Termes pas à pas Terme(s) construit(s).
- Sélectionnez les options allant de *lninc* à *réside* comme termes pas à pas.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Régression logistique multinomiale.

Récapitulatif des étapes

Figure 5-33
Récapitulatif des étapes

Nombre d'imputations	Modèle	Action	Effect(s)	Critères d'ajustement du modèle	Tests de sélection d'effets		
				-2 log vraisemblance	Khi-deux[a]	degrés de liberté	Signif
Données initiales	0	Saisi	Constante	1353.555			
	1	Saisi	ed	1260.972	92.583	12	.000
	2	Saisi	employ	1237.664	23.308	3	.000
	3	Saisi	marital	1229.808	7.856	3	.049
1	0	Saisi	Constante	2762.531	.		
	1	Saisi	ed	2608.189	154.342	12	.000
	2	Saisi	employ	2563.671	44.518	3	.000
	3	Saisi	reside	2549.200	14.470	3	.002
	4	Saisi	address	2541.050	8.151	3	.043
2	0	Saisi	Constante	2762.531	.		
	1	Saisi	ed	2603.940	158.591	12	.000
	2	Saisi	employ	2563.367	40.573	3	.000
	3	Saisi	marital	2545.743	17.624	3	.001
	4	Saisi	address	2536.532	9.211	3	.027
3	0	Saisi	Constante	2762.531	.		
	1	Saisi	ed	2600.074	162.457	12	.000
	2	Saisi	employ	2558.560	41.514	3	.000
	3	Saisi	marital	2546.062	12.499	3	.006
	4	Saisi	address	2536.348	9.714	3	.021
4	0	Saisi	Constante	2762.531	.		
	1	Saisi	ed	2601.616	160.915	12	.000
	2	Saisi	employ	2558.463	43.153	3	.000
	3	Saisi	marital	2543.747	14.716	3	.002
	4	Saisi	address	2533.341	10.406	3	.015
5	0	Saisi	Constante	2762.531	.		
	1	Saisi	ed	2604.773	157.759	12	.000
	2	Saisi	employ	2561.792	42.980	3	.000
	3	Saisi	marital	2549.096	12.696	3	.005

Méthode pas à pas : Entrée ascendante

a. La valeur de Khi-deux de l'entrée est basée sur le test de ratio de vraisemblance.

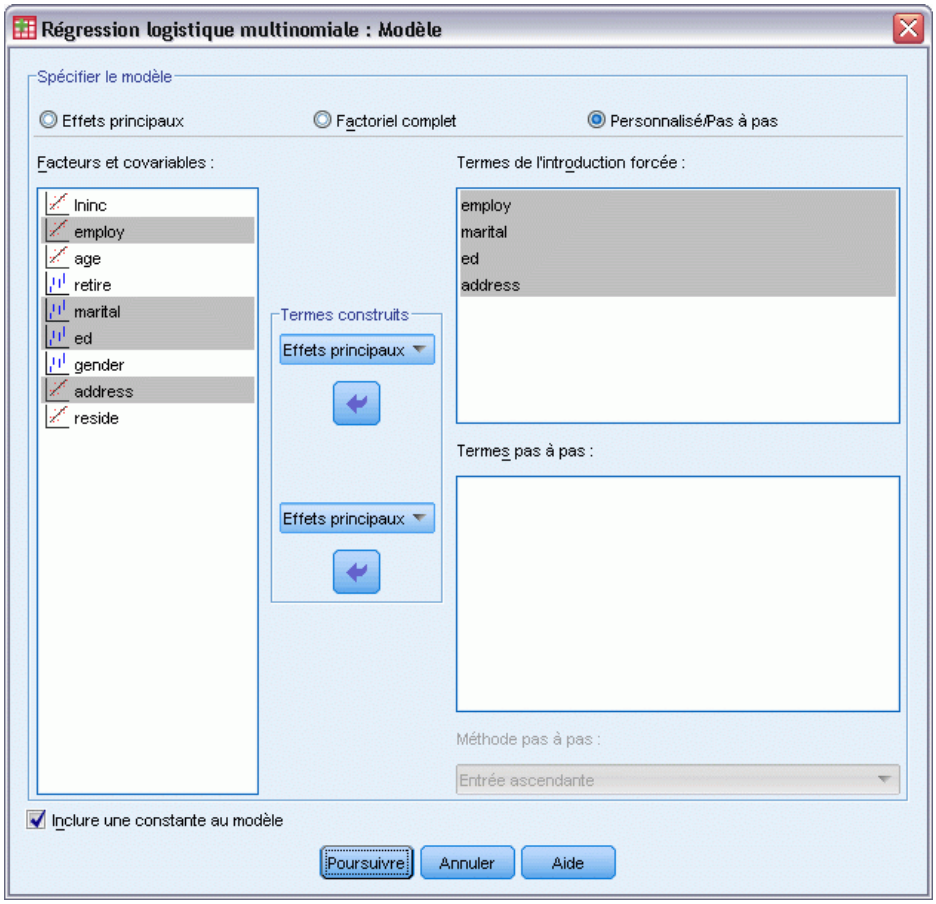
La régression logistique multinominale prend en charge le regroupement des coefficients de régression. Cependant, vous noterez que *tous* les tableaux des résultats présentent les résultats pour chaque imputation et pour les données d'origine. Ceci vient du fait que le fichier est scindé à *Imputation_*, et par conséquent tous les tableaux qui utilisent la variable de scission présenteront les groupes de fichiers scindés regroupés dans un seul tableau.

Vous observerez également que le tableau Estimations des paramètres ne présente pas d'estimations regroupées. Pour en connaître la raison, veuillez consulter le récapitulatif des étapes. Nous avons demandé une sélection pas à pas des effets de modèle et ce même ensemble d'effets n'a pas été choisi pour toutes les imputations. Par conséquent, le regroupement est impossible. Cependant, ceci fournit néanmoins des informations utiles car nous pouvons observer que *ed (Niveau d'éducation)*, *employ (Nb d'années avec l'employeur actuel)*, *marital (Situation familiae)* et *address (Nb d'années à la même adresse)* sont souvent choisies par la sélection pas à

pas parmi le imputations. Nous ajusterons un autre modèle en utilisant uniquement ces variables indépendantes.

Exécution du modèle avec un sous-ensemble de variables indépendantes

Figure 5-34
Boîte de dialogue Modèle

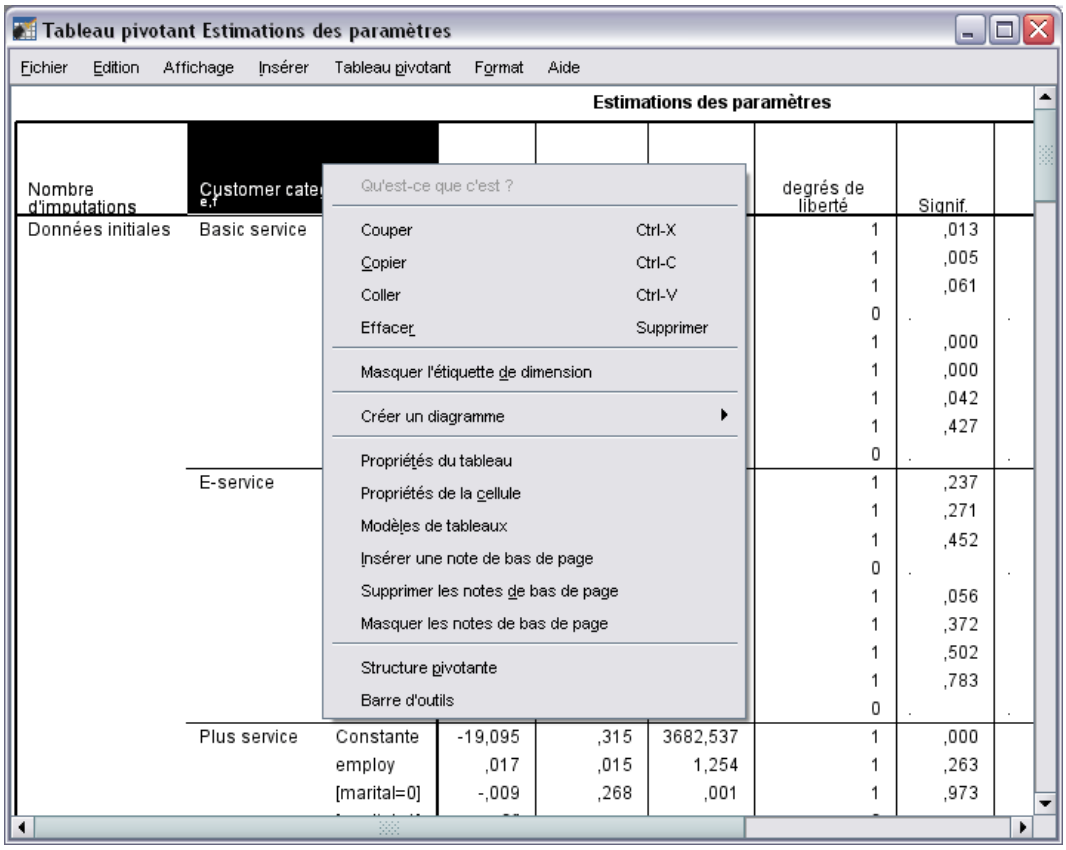


- Rappelez la boîte de dialogue Régression logistique multinomiale et cliquez sur Modèle.
- Désélectionnez les variables dans la liste Termes pas à pas.
- Sélectionnez Effets principaux dans la liste déroulante Termes de l'introduction forcée Terme(s) construit(s).
- Sélectionnez *employ*, *marital*, *ed* et *address* comme termes de l'introduction forcée.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Régression logistique multinomiale.

Estimations regroupées des paramètres

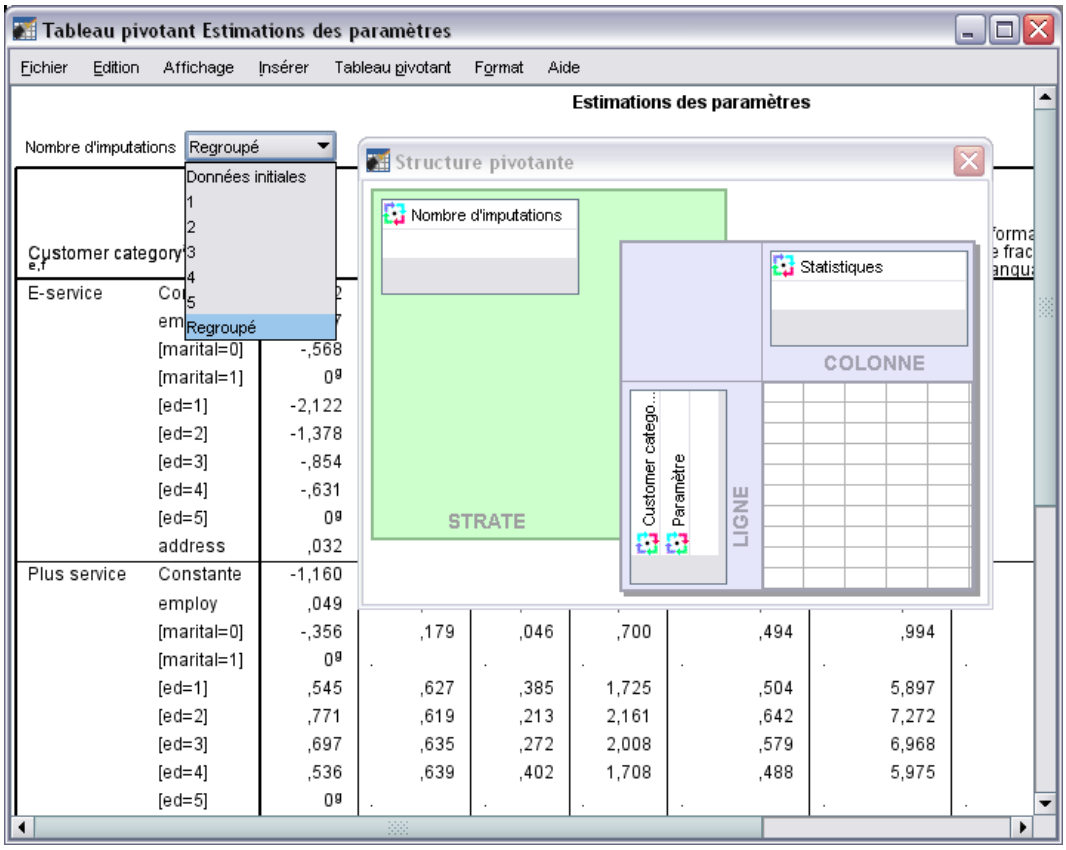
Ce tableau est relativement large, mais le faire pivoter offrira plusieurs vues différentes et utiles des résultats.

Figure 5-35
Estimations regroupées des paramètres



- Activez (double-cliquez sur) le tableau puis sélectionnez Structure pivotante dans le menu contextuel.

Figure 5-36
Estimations regroupées des paramètres



- Déplacez *Nombre d'imputations* de la ligne à la strate.
- Sélectionnez Regroupé dans la liste déroulante Nombre d'imputations.

Figure 5-37
Estimations regroupées des paramètres

Customer category[a,b,c,d,e,f]		B	Erreur std.	Signif.	Exp(B)	Intervalle de confiance 95% pour Exp(B)		Information de fraction manquante.	Variance d'augmentation relative	Efficacité relative
						Borne inférieure	Borne supérieure			
E-service	Constante	,622	,424	,142				,027	,028	,995
	employ	,027	,012	,020	1,028	1,004	1,052	,029	,030	,994
	[marital=0]	-,568	,196	,004	,566	,385	,833	,060	,062	,988
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-2,122	,474	,000	,120	,047	,304	,065	,067	,987
	[ed=2]	-1,378	,441	,002	,252	,106	,598	,033	,034	,993
	[ed=3]	-,854	,456	,062	,426	,174	1,043	,067	,069	,987
	[ed=4]	-,631	,443	,154	,532	,223	1,268	,016	,016	,997
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,032	,011	,005	1,033	1,010	1,056	,020	,020	,996
Plus service	Constante	-1,160	,610	,057				,014	,014	,997
	employ	,049	,011	,000	1,050	1,028	1,072	,033	,034	,993
	[marital=0]	-,356	,179	,046	,700	,494	,994	,012	,012	,998
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	,545	,627	,385	1,725	,504	5,897	,035	,036	,993
	[ed=2]	,771	,619	,213	2,161	,642	7,272	,026	,026	,995
	[ed=3]	,697	,635	,272	2,008	,579	6,968	,034	,035	,993
	[ed=4]	,536	,639	,402	1,708	,488	5,975	,026	,026	,995
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,020	,011	,075	1,020	,998	1,042	,095	,101	,981
Total service	Constante	1,084	,405	,007				,018	,018	,996
	employ	,039	,012	,001	1,039	1,015	1,064	,010	,010	,998
	[marital=0]	-,635	,194	,001	,530	,362	,776	,028	,028	,994
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-3,536	,534	,000	,029	,010	,083	,080	,083	,984
	[ed=2]	-1,768	,422	,000	,171	,075	,391	,018	,018	,996
	[ed=3]	-1,330	,437	,002	,264	,112	,622	,039	,040	,992
	[ed=4]	-,522	,420	,214	,593	,261	1,351	,015	,015	,997
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,017	,012	,179	1,017	,992	1,042	,092	,097	,982

Cette vue présente toutes les statistiques des résultats regroupés. Vous pouvez utiliser et interpréter ces coefficients comme vous utiliseriez ce tableau pour un ensemble de données sans valeurs manquantes.

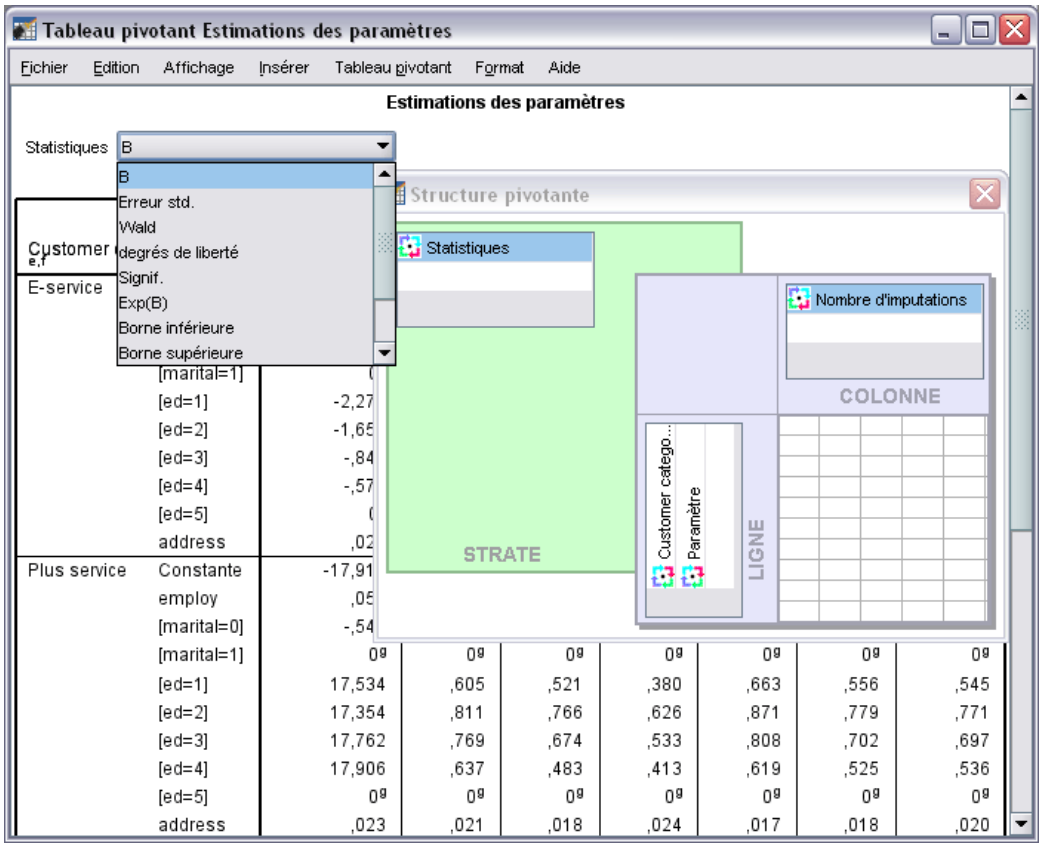
Le tableau des estimations de paramètres récapitule l'effet de chaque variable indépendante. Le rapport du coefficient à son erreur standard mis au carré égale la statistique de Wald. Si le seuil de signification de la statistique Wald est petit (inférieur à 0,05), alors le paramètre est différent de 0.

- Les paramètres avec des coefficients négatifs significatifs diminuent la vraisemblance de cette modalité de réponse par rapport à la modalité de référence.
- Les paramètres avec des coefficients positifs augmentent la vraisemblance de cette modalité de réponse.
- Les paramètres associés avec la dernière modalité de chaque facteur sont redondants selon la constante.

Le tableau contient trois colonnes supplémentaires qui offrent d'autres informations sur les résultats regroupés. La **fraction des informations manquantes** est une estimation du rapport entre les informations manquantes et les informations "complètes", basée sur l' **augmentation relative de la variance** provenant de la non-réponse qui, à son tour, est un rapport (modifié) de

la variance entre les imputations et la variance moyenne dans les imputations du coefficient de régression. L' **efficacité relative** est une comparaison de cette estimation avec une estimation (théorique) calculée à l'aide d'un nombre d'imputations infini. L'efficacité relative est déterminée par la fraction des informations manquantes et le nombre d'imputations utilisées pour obtenir le résultat regroupé. Lorsque la fraction des informations manquantes est importante, un nombre d'imputations plus élevé est nécessaire pour rapprocher l'efficacité relative de 1 et l'estimation regroupée de l'estimation idéale.

Figure 5-38
Estimations regroupées des paramètres



- A présent, réactivez (double-cliquez sur) le tableau puis sélectionnez Structure pivotante dans le menu contextuel.
- Déplacez *Nombre d'imputations* de la strate à la colonne.
- Déplacez *Statistiques* de la colonne à la strate.
- Sélectionnez B dans la liste déroulante Statistiques.

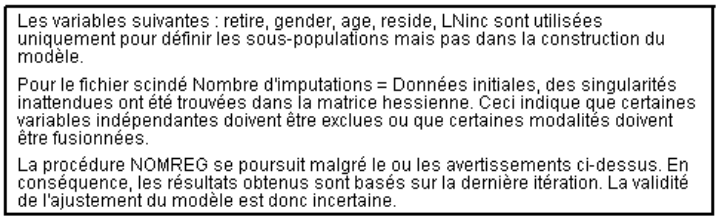
Figure 5-39
Estimations regroupées de paramètres, Nombre d'imputations dans les colonnes et Statistiques dans la strate

Statistiques=B

Customer category[a,b,c,d,e,f]		Nombre d'imputations						
		Données initiales	1	2	3	4	5	Regroupé
E-service	Constante	,637	,610	,701	,647	,527	,625	,622
	employ	,054	,027	,030	,028	,025	,027	,027
	[marital=0]	-,760	-,560	-,534	-,629	-,524	-,594	-,568
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	-2,272	-2,140	-2,284	-2,063	-1,991	-2,132	-2,122
	[ed=2]	-1,657	-1,403	-1,490	-1,359	-1,297	-1,343	-1,378
	[ed=3]	-,848	-,858	-1,019	-,827	-,724	-,841	-,854
	[ed=4]	-,576	-,636	-,705	-,618	-,562	-,636	-,631
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,028	,035	,032	,032	,031	,032	,032
Plus service	Constante	-17,912	-1,211	-1,150	-1,056	-1,220	-1,164	-1,160
	employ	,056	,048	,050	,049	,046	,050	,049
	[marital=0]	-,549	-,380	-,334	-,367	-,350	-,351	-,356
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	17,534	,605	,521	,380	,663	,556	,545
	[ed=2]	17,354	,811	,766	,626	,871	,779	,771
	[ed=3]	17,762	,769	,674	,533	,808	,702	,697
	[ed=4]	17,906	,637	,483	,413	,619	,525	,536
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,023	,021	,018	,024	,017	,018	,020
Total service	Constante	1,266	1,071	1,155	1,018	1,078	1,099	1,084
	employ	,044	,039	,040	,038	,038	,038	,039
	[marital=0]	-,522	-,627	-,684	-,635	-,608	-,621	-,635
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	-3,590	-3,666	-3,634	-3,400	-3,380	-3,599	-3,536
	[ed=2]	-2,133	-1,820	-1,812	-1,733	-1,700	-1,774	-1,768
	[ed=3]	-1,214	-1,269	-1,441	-1,244	-1,334	-1,365	-1,330
	[ed=4]	-,468	-,517	-,557	-,458	-,500	-,577	-,522
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,012	,018	,017	,019	,011	,019	,017

Cette vue du tableau est utile pour comparer les valeurs parmi les imputations, pour une visualisation rapide de la variation dans les estimations de coefficient de régression d'imputation à imputation et même par rapport aux données orginales. Plus spécifiquement, pour déplacer la statistique de la strate à l'écart-type. L'erreur vous permet d'observer la façon dont l'imputation multiple a réduit la variabilité dans les estimations de coefficients par rapport à l'élimination des observations incomplètes (données d'origine).

Figure 5-40
Avertissements


Les variables suivantes : retire, gender, age, reside, LNinc sont utilisées uniquement pour définir les sous-populations mais pas dans la construction du modèle.

Pour le fichier scindé Nombre d'imputations = Données initiales, des singularités inattendues ont été trouvées dans la matrice hessienne. Ceci indique que certaines variables indépendantes doivent être exclues ou que certaines modalités doivent être fusionnées.

La procédure NOMREG se poursuit malgré le ou les avertissements ci-dessus. En conséquence, les résultats obtenus sont basés sur la dernière itération. La validité de l'ajustement du modèle est donc incertaine.

Cependant, dans cet exemple, l'ensemble de données d'origine génère une erreur qui explique les estimations de paramètres particulièrement étendues pour la constante *Plus service* et les niveaux non-redondants de *ed (Niveau d'éducation)* dans la colonne des données d'origine du tableau.

Récapitulatif

A l'aide des procédures d'imputation multiple, vous avez analysé les modèles de valeurs manquantes et avez découvert que de nombreuses informations auraient certainement été perdues si l'élimination simple des observations incomplètes avait été utilisée. Après une exécution automatique initiale de l'imputation multiple, vous avez découvert que des contraintes étaient nécessaires pour conserver les valeurs imputées dans des limites raisonnables. L'exécution avec contraintes a produit des valeurs de qualité et il n'y a eu aucune preuve apparente que la méthode FCS n'a pas convergé. A l'aide de l'ensemble de données "complet" avec valeurs imputées, vous avez ajusté la régression logistique multinominale aux données et avez obtenu des estimations regroupées de régression. Vous avez également découvert que l'ajustement du modèle final n'aurait pas été possible avec l'élimination des observations incomplètes dans les données d'origine.

Annexe

A

Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire Echantillons pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espangol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

Descriptions

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.**Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **aflatoxin20.sav.** Ce fichier de données contient les mesures d'aflatoxine de chacun des 16 échantillons des récoltes 4 et 8 du fichier de données *aflatoxin.sav*.
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit

un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **autoaccidents.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un analyste en assurances pour modéliser le nombre d'accidents de la route par conducteur tout en prenant en compte l'âge et le sexe du conducteur. Chaque observation représente un conducteur distinct et enregistre son sexe, son âge et le nombre d'accidents de la route au cours des cinq dernières années.
- **band.sav.** Ce fichier de données contient les chiffres de ventes hebdomadaires hypothétiques de CD musicaux d'un groupe. Les données relatives à trois variables explicatives possibles sont également incluses.
- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondant aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique , on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique , on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= « aliment préféré » à 15= « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.

- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modelé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.
- **carpet.sav.** Dans un exemple courant , une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.
- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.

- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé . Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.
- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **customers_model.sav.** Ce fichier de données d'hypothèse concerne les personnes ciblées par une campagne de marketing. Ces données incluent des informations démographiques, un récapitulatif de l'historique d'achat et indiquent si chaque personne a répondu ou non à la campagne. Chaque observation représente une personne distincte.
- **customers_new.sav.** Ce fichier de données d'hypothèse concerne les personnes constituant des cibles potentielles pour une campagne de marketing. Ces données incluent des informations démographiques et un récapitulatif de l'historique d'achat pour chaque personne. Chaque observation représente une personne distincte.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.
- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.

- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.
- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman. Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui chercher à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.
- **guttman.sav.** Bell a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au

cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).

- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.
- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité 15×15, dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.
- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér*(ation) et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.

- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.**Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des Etats-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents , , ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.
- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **polishing.sav.** Il s'agit du fichier de données du « Nambeware Polishing Times » de la Data and Story Library. Il concerne les mesures qu'entreprend un fabricant de vaisselle en métal (Nambe Mills, Santa Fe, Nouveau-Mexique) pour planifier sa production. Chaque observation représente un article différent de la gamme de produits. Le diamètre, le temps de polissage, le prix et le type de produit sont enregistrés pour chaque article.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.

- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.
- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.**Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. A l'issue du cours de formation, chaque employé est testé et sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.
- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.

- **screws.sav.**Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous..
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. A intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».
- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (*http://dx.doi.org/10.3886/ICPSR02934*) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.
- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.

- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.
- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs .
- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs .
- **verd1985.sav.** Ce fichier de données concerne une enquête . Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'*âge* et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal*

domestique est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.

- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants . Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

Annexe

B

Notices

Licensed Materials – Property of SPSS Inc., an IBM Company. © Copyright SPSS Inc. 1989, 2010.

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Index