IBM SPSS Missing Values 28

Important

Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations générales figurant à la section «Remarques», à la page 21.

Notice d'édition

La présente édition s'applique à la version 28.0.0 d'IBM® SPSS Statistics et à toutes les éditions et modifications ultérieures sauf mention contraire dans les nouvelles éditions.

Table des matières

Chapitre 1. Valeurs manquantes

Les fonctions de valeurs manquantes suivantes sont incluses dans SPSS Statistics Premium Edition ou l'option Valeurs manquantes.

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 relatives à l'Imputation multiple proposent des analyses de motifs de données manquantes, orientées vers une imputation multiple finale des valeurs manquantes. C'est-à-dire que plusieurs versions du jeu de données sont produites, chacune d'elle contenant son propre jeu de données imputées. Lorsque des analyses statistiques sont effectuées, les estimations de paramètre pour tous les jeux de données imputés sont regroupées en pool 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 de 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 :

1. **Examiner le caractère manquant :** Utilisez l'analyse des valeurs manquantes et Analyser les motifs pour explorer des motifs de valeurs manquantes dans vos données et déterminer si l'imputation multiple est nécessaire.
2. **Inclure les valeurs manquantes :** Utilisez Imputer des valeurs de données manquantes pour imputer les valeurs manquantes.
3. **Analyser les données "complètes" :** Utilisez n'importe quelle procédure prenant en charge les données à imputation multiple. Pour plus d'informations sur l'analyse des jeux de données à imputation multiple et pour obtenir la liste des procédures prenant en charge ces données, voir «Analyse de données à imputation multiple», à la page 15.

Analyse des valeurs manquantes

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

- Elle décrit le motif 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 ?
- Elle estime les moyennes, les écarts types, les covariances et les corrélations pour différentes méthodes relatives aux valeur manquantes : Toute observation incomplète, Seulement composantes

non valides, 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 motif 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 : motifs 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 de l'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 de l'utilisateur.

Pondérations de fréquence : 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 motif 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 factorielle 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

1. A partir des menus, sélectionnez :

   Analyse > **Analyse des valeurs manquantes...**
2. 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 catégorielles (numériques ou chaîne) et entrer une limite relative au nombre de catégories (**Catégories maximales**).
- Cliquer sur **Motifs** pour mettre en tableau les motifs de données manquantes. Pour plus d'informations, voir «Affichage des motifs de valeurs manquantes», à la page 3.
- Cliquer sur **Descriptives** pour afficher les statistiques descriptives des valeurs manquantes. Pour plus d'informations, voir «Affichage des statistiques descriptives des valeurs manquantes», à la page 4.
- Sélectionner une méthode d'estimation des statistiques (moyennes, covariances et corrélations) et, éventuellement, d'imputation des valeurs manquantes. Pour plus d'informations, voir «Estimation des statistiques et imputation des valeurs manquantes», à la page 5.
- Si vous sélectionnez EM ou Régression, cliquez sur **Variables...** pour spécifier le sous-ensemble à utiliser pour l'estimation. Pour plus d'informations, voir «Variables dépendantes et de prédicteur», à la page 7.
- Sélectionnez une variable de libellé d'observation. Cette variable permet de libeller les observations dans les tableaux de motifs qui affichent des observations individuelles.

Affichage des motifs de valeurs manquantes

Vous pouvez choisir d'afficher différents tableaux montrant les motifs 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 motifs de données manquantes.

Observations mises en tableau : Les motifs de valeurs manquantes dans les variables d'analyse sont mis en tableau, avec affichage des fréquences pour chaque motif. Utilisez l'option **Trier les variables par motif de valeur manquante** pour indiquer si les effectifs et les variables sont triés selon la similarité des motifs. Utilisez l'option **Omettez les motifs avec moins de n % d'observation** pour éliminer les motifs 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 motif de valeur manquante** pour indiquer si les effectifs et les variables sont triés selon la similarité des motifs.

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 système manquante

A. Premier type de valeur manquante de l'utilisateur

B. Second type de valeur manquante de l'utilisateur

C. Troisième type de valeur manquante de l'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 motifs manquants. Pour les variables quantitatives (échelle), c'est la moyenne qui apparaît ; dans le cas des variables catégorielles, il s'agit du nombre d'observations correspondant à un motif dans chacune des catégories.

- **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 motifs de valeurs manquantes

1. Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez afficher les motifs de valeurs manquantes.
2. Cliquez sur **Motifs**.
3. Sélectionnez les tableaux de motif à afficher.

Affichage des statistiques descriptives des valeurs manquantes

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 indicateur

Pour chaque variable, une variable indicateur est créée. Cette variable catégorielle indique si la variable est présente ou manquante pour une observation individuelle. Les variables indicateur permettent de créer la disparité, le test *t* et les tables 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 de diagonale contient le pourcentage des valeurs manquantes pour une seule variable.

Tests t avec groupes formés d'après les variables indicateur : 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.

Tableaux croisés de variables indicateurs et qualitatives : Un tableau est affiché pour chaque variable catégorielle. Pour chacune des catégories, 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.

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

Pour afficher des statistiques descriptives

1. Dans la boîte de dialogue principale Analyse des valeurs manquantes, sélectionnez les variables pour lesquelles vous souhaitez afficher les statistiques descriptives des valeurs manquantes.
2. Cliquez sur **Descriptives**.
3. 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 de 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 de vraisemblance.

La statistique du khi-carré 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

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 rapport 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.

Enregistrer les données complétées : Vous pouvez enregistrer un jeu 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

1. 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.
2. Sélectionnez **EM** dans le groupe Estimation.
3. Pour spécifier les variables dépendantes (prévues) et de prédicteurs, cliquez sur **Variables**. Pour plus d'informations, voir «Variables dépendantes et de prédicteur», à la page 7.
4. Cliquez sur **EM**.
5. Sélectionnez les options EM de votre choix.

Options de l'estimation de la 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, variables aléatoires 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.
- *Variables aléatoires normales*. Les termes d'erreur sont choisis de façon aléatoire à partir d'une distribution contenant une valeur minimum attendue de 0 et un écart type égal à la racine carrée du terme d'erreur quadratique moyenne de la régression.
- *Variables aléatoires t de Student*. Les termes d'erreur sont choisis de manière aléatoire à partir de la distribution t, et redimensionnés par l'erreur quadratique de la moyenne des carrés (RMSE).

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

Enregistrer les données complétées : Ecrit un jeu 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

1. 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.
2. Sélectionnez **Régression** dans le groupe Estimation.
3. Pour spécifier les variables dépendantes (prévues) et de prédicteurs, cliquez sur **Variables**. Pour plus d'informations, voir «Variables dépendantes et de prédicteur», à la page 7.
4. Cliquez sur **Régression**.
5. Sélectionnez les options de régression de votre choix.

Variables dépendantes et de prédicteur

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 de prédicteur 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 classements des infirmières et un autre les classements 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 de prédicteur

1. 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.
2. Sélectionnez **EM** ou **Régression** dans le groupe Estimation.
3. Cliquez sur **Variables**.
4. Si vous souhaitez utiliser des variables spécifiques, plutôt que la totalité des variables, en guise de variables dépendantes et de variables de prédicteur, sélectionnez **Sélectionner les variables**, puis déplacez les variables vers les listes appropriées.

Commande MVA. Fonctions additionnelles

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Spécifier différentes variables descriptives pour les motifs de valeur manquante, les motifs de données et les motifs 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 motifs de données à l'aide de la sous-commande DPATTERN.
- Spécifier plusieurs variables de tri pour les motifs 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 au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

imputation multiple

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

Considérations sur les données à imputation multiple

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

- *Nominal*. Une variable peut être traitée comme étant nominale si ses valeurs représentent des catégories 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 catégories 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 catégories 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. Pour modifier le niveau de mesure d'une variable de manière permanente,

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

Tableau 1. Icônes de niveau de mesure

	Numérique	**Chaîne**	**Date**	**Heure**
Echelle (continue).		n/a		
Ordinal				
Champs nominaux				

Pondérations de fréquence : 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.

Echantillons complexes : La procédure d'Imputation multiple ne traite pas de manière explicite les strates, les clusters 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 procédures d'échantillonnage complexe n'analysent pas de manière automatique les jeux de données à imputation multiple. Pour obtenir la liste complète des procédures prenant en charge le regroupement en pool, voir «Analyse de données à imputation multiple», à la page 15.

Valeurs manquantes : Les valeurs manquantes utilisateur et système 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 prédicteurs dans les modèles d'imputation. Les valeurs manquantes utilisateur et système 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 motifs contient des mesures descriptives des motifs de valeurs manquantes dans les données et peut servir d'étape d'exploration avant l'imputation.
- Imputer les valeurs des données manquantes permet de générer des imputations multiples. Les jeux de données complets peuvent être analysés avec des procédures prenant en charge des jeux de données à imputation multiple. Pour plus d'informations sur l'analyse des jeux de données à imputation multiple et pour obtenir la liste des procédures prenant en charge ces données, voir «Analyse de données à imputation multiple», à la page 15.

Analyse des motifs

La fonction Analyser les motifs contient des mesures descriptives des motifs 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 motifs 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 motifs des valeurs manquantes peut contribuer à déterminer les étapes suivantes de l'imputation.

A partir des menus, sélectionnez :

Analyse > **Imputation multiple** > **Analyser les motifs...**

1. Sélectionnez au moins deux variables d'analyse. La procédure analyse les motifs 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.

Sortie : La sortie facultative suivante est disponible :

- **Récapitulatif des valeurs manquantes :** Il affiche un graphique circulaire de panels 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.
- **Motifs de valeurs manquantes :** Permet d'afficher des motifs mis en tableau de valeurs manquantes. Chaque motif correspond à un groupe d'observations avec le même motif de données complètes et incomplètes dans les variables d'analyse. Vous pouvez utiliser ces sorties 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 motif monotone. La procédure ordonne les variables d'analyse pour révéler ou ressembler à un motif monotone. Si aucun motif monotone n'existe après la réorganisation, vous pouvez en conclure que les données ont un motif 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, la sortie ne comprendra que 15 variables.

Imputation des valeurs de données manquantes

L'imputation des valeurs de données manquantes permet de générer des imputations multiples. Les jeux de données complets peuvent être analysés avec des procédures prenant en charge des jeux de données à imputation multiple. Pour plus d'informations sur l'analyse des jeux de données à imputation multiple et pour obtenir la liste des procédures prenant en charge ces données, voir «Analyse de données à imputation multiple», à la page 15.

Exemple : Un fournisseur de services de télécommunication souhaite mieux comprendre les motifs 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 le jeu de données.

A partir des menus, sélectionnez :

Analyse > **Imputation multiple** > **Imputer les valeurs des données manquantes...**

1. Sélectionnez au moins deux variables dans le modèle d'imputation. La procédure impute des valeurs multiples pour les données manquantes de ces variables.
2. Spécifiez le nombre d'imputations à calculer. Par défaut, cette valeur est 5.
3. Spécifiez un jeu de données ou un fichier de données au format IBM SPSS Statistics dans lequel les données imputées devront être écrites.

   Le jeu 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 le jeu de données initial comprend 100 observations et que vous ayez 5 imputations, le jeu de données de sortie comportera 600 observations. Toutes les variables dans le jeu de données d'entrée sont incluses dans le jeu de données de sortie. Les propriétés du dictionnaire (noms, libellés, etc.) des variables existantes sont copiées dans le nouveau jeu 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 du jeu de données de sortie. Si des scissions sont actives lorsque la procédure est exécutée, le jeu 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) du jeu 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.

Analyser les données : Lit les données dans le jeu de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si le jeu de données est important, cette action peut prendre un certain temps.

Affecter 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 la vue de variable 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

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

Méthode d'imputation

La méthode **Automatique** analyse les données et utilise la méthode monotone si les données présentent un motif 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 motif de données manquantes est arbitraire (monotone ou non).

Pour chaque itération et pour chaque variable dans l'ordre 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 prédicteurs, 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 le jeu 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 Sortie, vous pouvez enregistrer les données de l'historique des itérations FCS et les visualiser sous forme de tracé pour évaluer la convergence.

Monotone

Méthode non-itérative pouvant être utilisée uniquement lorsque les données présentent un motif de valeurs manquantes monotone. Un motif 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 motif 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 prédicteurs, et impute ensuite les valeurs manquantes pour la variable à ajuster. Ces valeurs imputées sont enregistrées dans le jeu de données imputé.

Inclure des interactions d'ordre 2
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 de prédicteur. Lorsqu'une méthode spécifique est choisie, vous pouvez, si vous le désirez, inclure toutes les interactions bidirectionnelles possibles parmi les variables de prédicteur catégoriel.

Type de modèle pour les variables d'échelle

Régression linéaire
Lorsque la méthode d'imputation est automatiquement sélectionnée, la régression linéaire est utilisée comme modèle univarié pour les variables d'échelle.

Egalisation par la moyenne prédictive (PMM)
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 garantit que les valeurs imputées sont plausibles. Pour PMM, la valeur imputée est basée sur la valeur définie pour la valeur **Sélectionner une observation complète de manière aléatoire à partir des prévisions les plus proches (k)**, où (*k*) est un entier positif dont la valeur par défaut est 5.

La régression logistique est toujours utilisée comme modèle univarié pour les variables catégorielles. Indépendamment du type de modèle, les prédicteurs catégoriels sont traités à l'aide de la codification par indicateurs (factice).

Tolérance de 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

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 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 prédicteurs. Généralement, chaque variable d'analyse est considérée à la fois comme une variable dépendante et comme un prédicteur dans le modèle d'imputation. Le **Rôle** peut servir à désactiver l'imputation pour les variables que vous souhaitez **Utiliser comme prédicteur uniquement** ou pour que des variables ne soient pas utilisées comme des prédicteurs (**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 prédicteurs.
- **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).

Tableau 2. Résultats d'arrondi

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

Exclure les variables avec de grandes quantités de données manquantes : Généralement, les variables d'analyse sont imputées et utilisées comme prédicteurs 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 prédicteurs 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.

Sortie

Affichage : Affichage des contrôles 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 prédicteurs 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 jeu de données contenant les données de l'historique des itérations pour l'imputation FCS. Le jeu 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 tracer les données sous forme de graphique pour mieux évaluer la convergence du modèle.

Commande IMPUTATION MULTIPLE - Fonctions additionnelles

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Spécifier un sous-ensemble de variables dont les statistiques descriptives sont affichées (sous-commandeRECAPITULATIFSIMPUTATIONS).
- Spécifier à la fois une analyse de motifs manquants et de l'imputation en n'exécutant la procédure qu'une seule fois.
- Spécifier le nombre maximal de paramètres de modèle autorisé lors de l'imputation d'une variable (mot-cléMAXMODELPARAM).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Utilisation des données à imputation multiple

Lorsqu'un jeu de données à imputation multiple (IM) est créé, une variable appelée *Imputation_* avec un libellé de variable *Nombre d'imputations* est ajoutée et le jeu de données est trié dans l'ordre croissant. Les observations du jeu 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 jeu de données, la présence de la variable *Imputation_* identifie le jeu de données comme un jeu de données IM possible.

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

Le jeu 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 jeu 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...**

1. Sélectionnez **Comparer les groupes**.
2. 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). Lorsque vous créez un nouveau jeu 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.

Pour activer le marquage, dans les menus de l'éditeur de données, choisissez :

Affichage > **Marquer les données imputées...**

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 Vue de données de l'éditeur de données.

Déplacement entre les imputations

1. A partir des menus, sélectionnez :

   Edition > **Aller à l'imputation...**
2. Sélectionnez l'imputation (ou données d'origine) dans la liste déroulante proposée.

Vous pouvez également sélectionner l'imputation dans la liste déroulante de la barre d'édition dans Vue de données de l'Editeur de données.

La position relative des observations est conservée lors de la sélection des imputations. Par exemple, si le jeu de données initial contient 1000 observations, l'observation 1034, la 34e observation de la première imputation, apparaît en haut de la grille. Si vous sélectionnez l'imputation **2** dans la liste déroulante, l'observation 2034, 34e 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 en pool de résultats d'une analyse de jeux 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 en pool. 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 en pool, contrairement aux procédures Rapport, Tracés P-P et Tracés Q-Q .

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

- **Regroupement en pool des tableaux de sorties :** Par défaut, lorsque vous exécutez une procédure prise en charge dans un jeu 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 en pool (finaux) qui prennent en compte les variations entre les imputations. Les statistiques qui sont regroupées en pool varient selon la procédure.
- **Regroupement en pool de PMML :** Vous pouvez également obtenir des PMML regroupés en pool à partir des procédures prises en charge qui exportent les PMML. Le PMML regroupé en pool est demandé de la même façon que le PMML non regroupé en pool (qu'il remplace lorsqu'il est enregistré).

Les procédures non prises en charge ne produisent ni sorties regroupées en pool ni fichiers PMML regroupés en pool.

Niveaux de regroupement en pool

Les sorties sont regroupées en pool à l'un des deux niveaux suivants :

- **Combinaison naïve :** Seul le paramètre regroupé en pool est disponible.
- **Combinaison univariée :** Le paramètre regroupé en pool, 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

en pool (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 regroupés en pool. Lorsque l'erreur standard d'une statistique est disponible, le regroupement en pool univarié est alors utilisé. Autrement, c'est le regroupement naïf qui est utilisé.

Procédures prenant en charge le regroupement en pool

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

Effectifs : Les fonctions suivantes sont prises en charge :

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

Descriptives : Les fonctions suivantes sont prises en charge :

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

Tableaux croisés : Les fonctions suivantes sont prises en charge :

- Le tableau croisé prend en charge les effectifs en regroupement en pool naïf.

Moyennes : Les fonctions suivantes sont prises en charge :

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

Test T pour échantillon unique : Les fonctions suivantes sont prises en charge :

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

Test T pour échantillons indépendants : Les fonctions suivantes sont prises en charge :

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

Test T pour échantillons appariés : Les fonctions suivantes sont prises en charge :

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

ANOVA à 1 facteur : Les fonctions suivantes sont prises en charge :

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

GLM - Univarié : Les fonctions suivantes sont prises en charge :

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

Modèles mixtes linéaires : Les fonctions suivantes sont prises en charge :

- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement en pool naïf.
- Le tableau Estimations des effets fixes prend en charge les estimations en regroupement en pool univarié.

- Le tableau Estimations des paramètres de covariance prend en charge les estimations en regroupement en pool univarié.
- Moyennes marginales estimées : Le tableau Estimations prend en charge la moyenne en regroupement en pool univarié.
- Moyennes marginales estimées : Le tableau Comparaisons appariée prend en charge la différence moyenne en regroupement en pool univarié.

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

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

Corrélations bivariées : Les fonctions suivantes sont prises en charge :

- Le tableau Statistiques descriptives prend en charge la moyenne et N en regroupement en pool naïf.
- Le tableau Corrélations prend en charge les corrélations et N en regroupement en pool Univarié. Veuillez noter que les corrélations sont transformées à l'aide de la transformation *z* de Fisher avant le regroupement en pool puis retransformées après le regroupement en pool.

Corrélations partielles : Les fonctions suivantes sont prises en charge :

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

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

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

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

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

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

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

Régression ordinale : Les fonctions suivantes sont prises en charge :

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

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

- Le tableau Statistiques de groupes prend en charge la moyenne et N Valide en regroupement en pool naïf.
- Le tableau Matrices intragroupes combinés prend en charge les corrélations en regroupement en pool naïf.
- Le tableau Coefficients de la fonction discriminante canonique prend en charge les coefficients non standardisés en regroupement en pool naïf.
- Le tableau Fonctions aux centroïdes des groupes prend en charge les coefficients non standardisés en regroupement en pool naïf.
- Le tableau Coefficients de la fonction de classement prend en charge les coefficients en regroupement en pool naïf.

Test du khi-carré : Les fonctions suivantes sont prises en charge :

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

Test binomial : Les fonctions suivantes sont prises en charge :

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

Suites en séquences : Les fonctions suivantes sont prises en charge :

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

Test Kolmogorov-Smirnov pour un échantillon : Les fonctions suivantes sont prises en charge :

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

Tests pour deux échantillons indépendants : Les fonctions suivantes sont prises en charge :

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

Tests pour plusieurs échantillons indépendants : Les fonctions suivantes sont prises en charge :

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

Tests pour deux échantillons liés : Les fonctions suivantes sont prises en charge :

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

Tests pour plusieurs échantillons liés : Les fonctions suivantes sont prises en charge :

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

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

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

Options d'imputation multiple

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

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 jeux 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.

Sortie d'analyse : Ce groupe contrôle le type de sortie du visualiseur produits lorsqu'un jeu de données à imputation multiple est analysé. Par défaut, les sorties seront produites pour le jeu de données d'origine (pré-imputation) et pour chacun des jeux de données imputés. De plus, pour ce genre de procédures qui prennent en charge le regroupement en pool de données imputées, des résultats finaux regroupés en pool seront générés. Lorsqu'un regroupement en pool univarié sera effectué, les diagnostiques de regroupement en pool seront également affichés. Mais vous pouvez supprimer toutes les sorties que vous ne désirez pas voir.

Pour définir les options d'imputation multiple

A partir des menus, sélectionnez :

Edition > **Options**

Cliquez sur l'onglet Imputation multiple.

Remarques

Le présent document a été développé pour des produits et des services proposés aux Etats-Unis et peut être mis à disposition par IBM dans d'autres langues. Toutefois, il peut être nécessaire de posséder une copie du produit ou de la version du produit dans cette langue pour pouvoir y accéder.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, programme ou service IBM n'implique pas que seul ce produit, programme ou service IBM puisse être utilisé. Tout autre élément fonctionnellement équivalent peut être utilisé, s'il n'enfreint aucun droit d'IBM. Il est de la responsabilité de l'utilisateur d'évaluer et de vérifier lui-même les installations et applications réalisées avec des produits, logiciels ou services non expressément référencés par IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. La remise de ce document ne vous donne aucun droit de licence sur ces brevets ou demandes de brevet. Si vous désirez recevoir des informations concernant l'acquisition de licences, veuillez en faire la demande par écrit à l'adresse suivante :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A

Pour toute demande d'informations relatives au jeu de caractères codé sur deux octets, contactez le service de propriété intellectuelle IBM ou envoyez vos questions par courrier à l'adresse suivante :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFAÇON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE. Certaines juridictions n'autorisent pas l'exclusion des garanties implicites, auquel cas l'exclusion ci-dessus ne vous sera pas applicable.

Le présent document peut contenir des inexactitudes ou des coquilles. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation de sa part, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci), et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Ces informations peuvent être soumises à des conditions particulières, prévoyant notamment le paiement d'une redevance.

Le logiciel sous licence décrit dans ce document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du Livret Contractuel IBM, des Conditions Internationales d'Utilisation de Logiciels IBM, des Conditions d'Utilisation du Code Machine ou de tout autre contrat équivalent.

Les données de performances et les exemples de clients sont fournis à titre d'exemple uniquement. Les performances réelles peuvent varier en fonction des configurations et des conditions d'exploitation spécifiques.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée uniquement comme un objectif.

Le présent document peut contenir des exemples de données et de rapports utilisés couramment dans l'environnement professionnel. Ces exemples mentionnent des noms fictifs de personnes, de sociétés, de marques ou de produits à des fins illustratives ou explicatives uniquement. Toute ressemblance avec des noms de personnes, de sociétés ou des données réelles serait purement fortuite.

LICENCE DE COPYRIGHT :

Le présent logiciel contient des exemples de programmes d'application en langage source destinés à illustrer les techniques de programmation sur différentes plateformes d'exploitation. Vous avez le droit de copier, de modifier et de distribuer ces programmes exemples sous quelque forme que ce soit et sans paiement d'aucune redevance à IBM, à des fins de développement, d'utilisation, de vente ou de distribution de programmes d'application conformes aux interfaces de programmation des plateformes pour lesquels ils ont été écrits ou aux interfaces de programmation IBM. Ces programmes exemples n'ont pas été rigoureusement testés dans toutes les conditions. Par conséquent, IBM ne peut garantir expressément ou implicitement la fiabilité, la maintenabilité ou le fonctionnement de ces programmes. Les programmes exemples sont fournis "EN L'ETAT", sans garantie d'aucune sorte. IBM ne sera en aucun cas responsable des dommages liés à l'utilisation des programmes exemples.

Toute copie totale ou partielle de ces programmes exemples et des oeuvres qui en sont dérivées doit comprendre une notice de copyright, libellée comme suit :

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines Corp. dans de nombreux pays. Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information" à l'adresse www.ibm.com/legal/copytrade.shtml.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques d'Adobe Systems Incorporated aux Etats-Unis et/ou dans certains autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium et Pentium sont des marques d'Intel Corporation ou de ses filiales aux Etats-Unis et/ou dans certains autres pays.

Linux est une marque de Linus Torvalds aux Etats-Unis et/ou dans certains autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques de Microsoft Corporation aux Etats-Unis et/ou dans certains autres pays.

UNIX est une marque enregistrée de The Open Group aux Etats-Unis et/ou dans certains autres pays.

Java ainsi que toutes les marques et tous les logos incluant Java sont des marques d'Oracle et/ou de ses sociétés affiliées.

Index