IBM SPSS Regression 19

Note: Before using this information and the product it supports, read the general information under Notices sur p. 46.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Régression fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Régression 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

Chapitre
1

Choix d'une procédure pour la Régression logistique binaire

Les modèles de régression logistique binaire peuvent être ajustés au moyen de la procédure de régression logistique ou de la procédure de régression logistique multinomiale. Chacune de ces procédures comporte des options qui lui sont propres. Il convient de faire une importante distinction théorique entre les deux procédures : la procédure de régression logistique génère toutes les prévisions, résidus, statistiques d'influence et tests de qualité d'ajustement en utilisant les données au niveau des observations individuelles, quelle que soit la façon dont ces données ont été entrées et que le nombre de paramètres de covariable soit inférieur ou non au nombre total d'observations ; alors que la procédure de régression logistique multinomiale agrège les observations au niveau interne pour constituer des sous-populations présentant des paramètres de covariable identiques pour les variables indépendantes, générant ainsi des prévisions, des résidus et des tests de qualité d'ajustement en fonction de ces sous-populations. Si toutes les variables indépendantes sont qualitatives ou que des variables indépendantes continues prennent en compte un nombre limité de valeurs—de sorte qu'il existe plusieurs observations pour chaque type de covariable distinct—, la méthode de constitution de sous-populations peut générer des tests de qualité d'ajustement et des résidus informatifs valides, ce qui n'est pas le cas de la procédure effectuée au niveau des observations individuelles.

La **régression logistique** offre les fonctionnalités spécifiques suivantes :

- test Hosmer-Lemeshow de la qualité d'ajustement du modèle ;
- analyses pas à pas ;
- contrastes permettant de définir le paramétrage du modèle ;
- césures alternatives pour le classement ;
- Diagrammes de classement
- modèle ajusté sur un ensemble d'observations par rapport à un ensemble d'observations présenté ;
- enregistrement des prévisions, des résidus et des statistiques d'influence.

La **régression logistique multinomiale** offre les fonctionnalités spécifiques suivantes :

- tests du Khi-deux de Pearson et de déviance pour la qualité d'ajustement du modèle ;
- définition de sous-populations pour le regroupement de données afin d'effectuer des tests de qualité d'ajustement ;
- énumération des effectifs, des effectifs prédits et des résidus par sous-population ;
- correction des estimations de variance pour la surdispersion

- matrice de covariance des estimations de paramètres ;
- tests des combinaisons linéaires de paramètres ;
- définition explicite des modèles emboîtés ;
- ajustement 1-1 de modèles de régression logistique conditionnels correspondants au moyen de variables différenciées.

Chapitre
2

Régression logistique

La régression logistique est utile lorsque vous souhaitez être capable de prévoir la présence ou l'absence d'une caractéristique ou d'un résultat en fonction de certaines valeurs ou d'un groupe de variables prédites. Elle est similaire à la régression linéaire mais elle convient aux modèles dans lesquelles les variables sont dichotomiques. Les coefficients de la régression logistique peuvent servir à estimer des odds ratios pour chacune des variables indépendantes d'un modèle. La régression logistique s'applique à une plus large gamme de situations de recherche que l'analyse discriminante.

Exemple : Quelles sont les caractéristiques du mode de vie qui constituent des facteurs de risques coronariens ? Sur un échantillon de patients choisis en fonction de leur statut de fumeur, leur régime alimentaire, leur consommation d'alcool et leur historique cardiaque, vous pouvez construire un modèle à l'aide de quatre variables du mode de vie pour expliquer la présence ou l'absence de déficiences coronariennes sur l'échantillon de patients. Le modèle peut alors servir à dériver les prévisions des odds ratios pour chaque facteur afin de vous indiquer, par exemple, que les fumeurs sont plus susceptibles de développer des déficiences coronariennes que les non-fumeurs.

Statistiques : Pour chaque analyse : observations totales, observations sélectionnées, observations valides. Pour chaque variable qualitative : codage du paramètre. Pour chaque pas : variable(s) saisie(s) ou supprimée(s), historique d'itération, –2 log-vraisemblance, qualité de l'ajustement, qualité d'ajustement de Hosmer-Lemeshow, Khi-deux, Khi-deux d'amélioration, tableau de classification, corrélations entre variables, groupes observés et diagramme des probabilités prévues, Khi-deux résiduel. Pour chaque variable de l'équation : coefficient (*B*), erreur standard *B*, statistique de Wald, odds ratio estimé (exp(*B*)), intervalle de confiance pour exp(*B*), log-vraisemblance si un terme a été supprimé du modèle. Pour chaque variable hors de l'équation : Statistique de significativité Pour chaque observation : groupe observé, probabilité prévue, groupe théorique, résidu, résidu standard.

Méthodes. Vous pouvez estimer des modèles à l'aide des entrées en bloc de variables ou de n'importe laquelle des méthodes par étapes suivantes : ascendante conditionnelle, ascendante rapport de vraisemblance, ascendante Wald, descendante conditionnelle, descendante rapport de vraisemblance, descendante Wald.

Données. Les variables dépendantes et indépendantes doivent être dichotomiques. Les variables indépendantes peuvent être de niveaux d'intervalles ou des variables qualitatives. Dans ce dernier cas, elles doivent être factices ou codées numériquement (il existe une option dans la procédure pour recoder les variables qualitatives automatiquement).

Hypothèses : La régression logistique ne s'appuie pas sur des hypothèses de distribution au même sens que l'analyse discriminante. Cependant, votre solution peut être plus stable si vos variables prédites suivent une distribution multivariée gaussienne. De surcroît, comme avec les autres

formes de régression, la multicolinéarité parmi les variables prédites peut entraîner une altération des estimations et l'augmentation des erreurs standard. La procédure est plus efficace lorsque l'appartenance au groupe est une variable purement qualitative, si l'appartenance au groupe est fondée sur des valeurs d'une variable continue(par exemple "QI élevé" opposé à "QI faible"), vous devez envisager d'utiliser la régression linéaire pour profiter de la richesse des informations offertes par la variable continue elle-même.

Procédures apparentées : Utilisez le diagramme de dispersion pour étudier la multicolinéarité de vos données. Si les hypothèses de normalité multivariées et d'égalité des matrices de variance/covariance sont satisfaites, vous devez obtenir une solution plus rapide à l'aide de la procédure d'analyse discriminante. Si toutes vos variables prédites sont qualitatives, vous pouvez également utiliser la procédure log-linéaire. Si votre variable dépendante est continue, utilisez la procédure de régression linéaire. Vous pouvez utiliser la procédure Courbe ROC pour représenter sous forme diagramme les probabilités enregistrées avec la procédure Régression logistique.

Obtenir une analyse de la régression logistique

▶ A partir des menus, sélectionnez :
Analyse > Régression > Logistique binaire...

Figure 2-1
Boîte de dialogue Régression logistique

▶ Sélectionnez une Variable dépendante dichotomique. Il peut s'agir d'une variable numérique ou d'une chaîne.

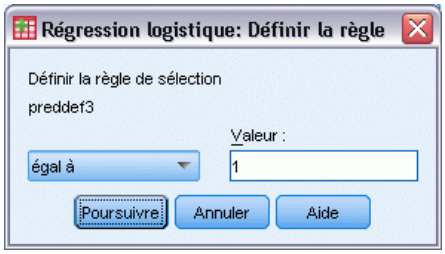
▶ Sélectionnez une ou plusieurs covariables. Pour ajouter des termes d'interaction, sélectionnez toutes les variables impliquées dans l'interaction, puis sélectionnez >a*b>.

Pour saisir les variables en groupe (**blocs**), sélectionnez les covariables pour un bloc, puis cliquez sur Suivant pour spécifier un nouveau bloc. Répétez jusqu'à ce que tous les blocs soient spécifiés.

Vous pouvez éventuellement sélectionner des observations pour analyse. Choisissez une variable de sélection, puis cliquez sur Loi.

Définir la règle de régression logistique

Figure 2-2
Boîte de dialogue Définir la règle de régression logistique



Les observations définies par la règle de sélection sont incluses dans l'estimation du modèle. Par exemple, si vous avez sélectionné une variable ainsi que l'opérateur égal à et que vous avez spécifié la valeur 5, seules les observations pour lesquelles la variable sélectionnée a une valeur égale à 5 sont incluses dans l'estimation du modèle.

Les résultats des statistiques et de classification sont générés pour les observations sélectionnées et celles qui ne le sont pas. Cette procédure met en œuvre un mécanisme de classification des nouvelles observations à partir des données précédemment existantes, ou de partitionnement de vos données en sous-ensembles de formation et de test, afin d'effectuer la validation du modèle généré.

Méthodes de sélection des variables de régression logistique

La sélection d'une méthode vous permet de spécifier la manière dont les variables indépendantes sont entrées dans l'analyse. En utilisant différentes méthodes, vous pouvez construire divers modèles de régression à partir du même groupe de variables.

- **Introduire.** Procédure de sélection de variables au cours de laquelle toutes les variables d'un bloc sont introduites en une seule opération.
- **Ascendante : conditionnelle.** Méthode de sélection pas à pas avec test d'entrée fondé sur la signification de la statistique de significativité et avec test de suppression fondé sur la probabilité d'une statistique du rapport de vraisemblance s'appuyant sur des estimations de paramètres conditionnels.
- **Ascendante : rapport de vraisemblance.** Méthode de sélection pas à pas avec test d'entrée fondé sur la signification de la statistique de significativité et avec test de suppression fondé sur la probabilité d'une statistique du rapport de vraisemblance s'appuyant sur des estimations de vraisemblance partielle maximale.
- **Ascendante : Wald.** Méthode de sélection pas à pas avec test d'entrée fondé sur la signification de la statistique de significativité et avec test de suppression fondé sur la probabilité de la statistique de Wald.
- **Descendante : conditionnelle.** Sélection progressive descendante. Le test de suppression se base sur la probabilité du rapport de vraisemblance calculé à partir d'estimations de paramètres conditionnels.

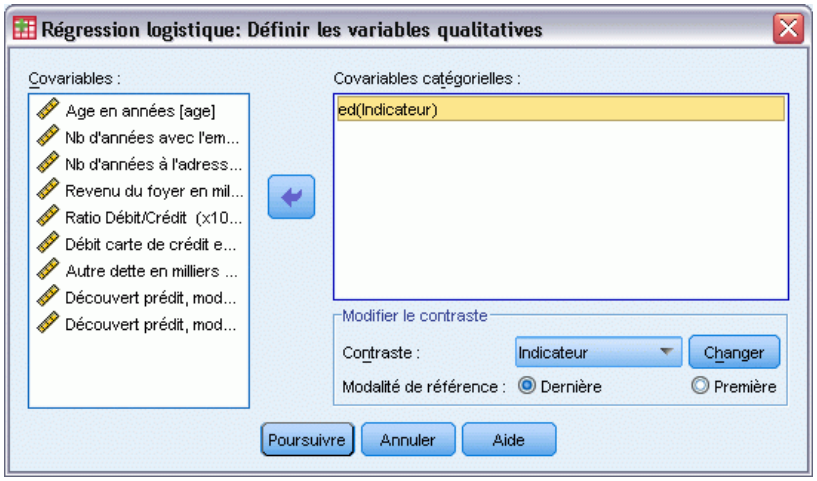
- **Descendante : rapport de vraisemblance (LR).** Sélection progressive descendante. Le test de suppression se base sur la probabilité de la statistique du rapport de vraisemblance calculé à partir des estimations de vraisemblance partielle maximale.
- **Descendante : Wald.** Sélection progressive descendante. Le test de suppression se base sur la probabilité de la statistique de Wald.

Les valeurs de significativité dans vos résultats sont basées sur l'adéquation à un modèle unique. Par conséquent, les valeurs de significativité ne sont généralement pas valides lorsqu'une méthode pas à pas est utilisée.

Toutes les variables indépendantes sélectionnées sont ajoutées dans un seul modèle de régression. Cependant, vous pouvez spécifier différentes méthodes d'entrée pour les sous-groupes de variables. Par exemple, vous pouvez entrer un bloc de variables dans le modèle de régression en utilisant la sélection pas à pas, et un second bloc en utilisant la sélection ascendante. Pour ajouter un second bloc de variables au modèle de régression, cliquez sur Suivant.

Régression logistique : définition des variables qualitatives

Figure 2-3
Boîte de dialogue Régression logistique : Variables qualitatives



Vous pouvez spécifier les détails de la manière dont la procédure de régression logistique gère les variables qualitatives :

Covariables : Contient la liste de toutes les covariables spécifiées dans la boîte de dialogue principale, soit par elles-mêmes, soit comme partie d'une interaction, à n'importe quelle strate. Si certaines de ces covariables sont des variables chaîne, vous pouvez utiliser des covariables qualitatives.

Covariables qualitatives : Etablit la liste de toutes les variables identifiées comme étant qualitatives. Chaque variable comprend une notation entre parenthèses indiquant le codage de contraste à utiliser. Les variables chaîne (identifiées par le symbole < suivi de leurs noms) sont déjà présentes dans la liste des covariables qualitatives. Sélectionnez n'importe quelle autre covariable qualitative à partir de la liste des covariables qualitatives.

Modifier le contraste : Permet de modifier la méthode de contraste. Les méthodes de contraste disponibles sont :

- **Indicateur :** Les contrastes indiquent la présence ou l'absence d'appartenance à la modalité. La modalité de référence est représentée par la matrice de contraste sous la forme d'une ligne de zéros.
- **Simple :** Chaque modalité de la variable prédite (hormis la modalité de référence) est comparée à la modalité de référence.
- **Différence :** Chaque modalité de la variable prédite (hormis la première modalité) est comparée avec l'effet moyen des modalités précédentes. (Aussi connu sous le nom de contrastes inversés d'Helmert.)
- **Helmert :** Chaque modalité de la variable prédite (hormis la dernière modalité) est comparée avec l'effet moyen des modalités suivantes.
- **Répété :** Chaque modalité de la variable prédite (hormis la première modalité) est comparée avec la modalité précédente.
- **Modèle polynomial :** Contraste polynomial orthogonal. On part de l'hypothèse que les modalités sont espacées de manière équivalente. Les contrastes polynomiaux sont utilisables pour les variables numériques seulement.
- **Ecart.** Chaque modalité de la variable prédite (hormis la modalité de référence) est comparée à l'effet global.

Si vous sélectionnez Ecart, Simple ou Indicateur, sélectionnez Première ou Dernière comme modalité de référence. Remarquez que vous ne changez pas réellement de méthode avant de cliquer sur Changer.

Les covariables chaîne doivent impérativement être des covariables qualitatives. Pour supprimer une variable chaîne de la liste des covariables qualitatives, vous devez supprimer tous les termes contenant cette variable de la liste des covariables de la boîte de dialogue principale.

Régression logistique : enregistrer les nouvelles variables

Figure 2-4
Boîte de dialogue Régression logistique : Enregistrer les nouvelles variables

Vous pouvez enregistrer les résultats de la régression logistique sous forme de nouvelles variables dans l'ensemble de données actif :

Prévisions : Enregistre les valeurs prévues par le modèle. Les options disponibles sont Probabilités et Groupe d'affectation.

- **Probabilités.** Enregistre pour chaque observation la probabilité d'occurrence prévue pour l'événement. Dans les résultats, un tableau affiche le nom et le contenu de toutes les nouvelles variables.
- **Groupes d'affectation prévus.** Le groupe qui possède la probabilité a posteriori la plus élevée, basé sur les écarts discriminants. Le groupe prévu par le modèle est celui auquel appartient l'observation.

Influence. Enregistre des valeurs à partir des statistiques qui mesurent l'influence des observations sur les prévisions. Les options disponibles sont Statistique de Cook, Bras de levier et Différence de bêta.

- **Statistique de Cook.** Régression logistique analogue à la statistique d'influence de Cook. Mesure permettant de savoir de combien les résidus de toutes les observations seraient modifiés si une observation donnée était exclue du calcul des coefficients de régression.
- **Valeur influente (ou bras de levier).** Influence relative de chaque observation sur la qualité d'ajustement du modèle.
- **Différence de bêta.** La différence de bêta correspond au changement des coefficients de régression qui résulte du retrait d'une observation particulière. Une valeur est calculée pour chaque terme du modèle, y compris la constante.

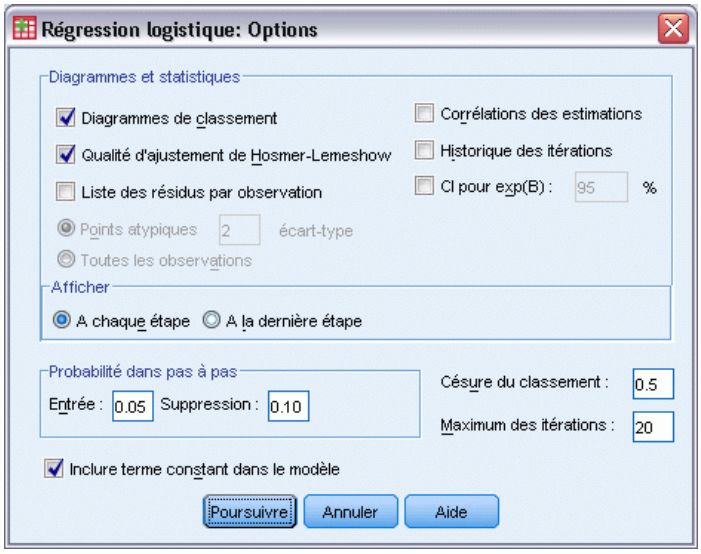
Résidus : Enregistre les résidus. Les options disponibles sont Non standardisés, Logit, Studentisés, Standardisés et Déviance.

- **Résidu non standardisé.** Différence entre la valeur observée et la valeur prévue par le modèle.
- **Résidu Logit.** Résidu de l'observation lorsque celle-ci est prévue dans l'échelle logit. Le résidu logit est le résidu divisé par la probabilité prévue fois 1, moins la probabilité prévue.
- **Résidu studentisé.** Evolution de la déviance du modèle lorsqu'une observation est exclue.
- **Résidus standardisés.** Résidu, divisé par une estimation de son erreur standard. Egalement appelés résidus de Pearson, les résidus standardisés ont une moyenne de 0 et un écart-type de 1.
- **Déviance.** Résidus fondés sur la déviance du modèle.

Exporter les informations du modèle dans un fichier XML : Les estimations de paramètres et leurs covariances (facultatif) sont exportées vers le fichier spécifié au format XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.

Options de régression logistique

Figure 2-5
Boîte de dialogue Régression logistique : Options



Vous pouvez sélectionner les options suivantes pour votre analyse :

Diagrammes et statistiques. Vous permet de demander statistiques et diagrammes. Les options disponibles sont Diagrammes de classement, Qualité d'ajustement d'Hosmer-Lemeshow, Liste des résidus par observation, Corrélations des estimations, Historique des itérations et CI pour exp(*B*). Sélectionnez l'une des options dans le groupe Affichage pour consulter les statistiques et les diagrammes soit A chaque pas, soit uniquement pour le modèle final, Au dernier pas.

- **Qualité d'ajustement de Hosmer-Lemeshow.** Cette statistique de qualité d'ajustement est plus robuste que la statistique de qualité d'ajustement traditionnellement utilisée pour la régression logistique, particulièrement pour les modèles ayant des covariables continues et les études d'échantillons de petite taille. Elle est basée sur le regroupement des observations en déciles de risque et la comparaison de la probabilité observée avec la probabilité théorique à l'intérieur de chaque décile.

Probabilité dans pas à pas : Vous permet de commander les critères d'insertion ou de suppression des variables dans l'équation. Vous pouvez spécifier les critères d'insertion ou de suppression des variables.

- **Probabilité pour méthode pas à pas.** Une variable est ajoutée au modèle si la probabilité de sa statistique de coordonnées principales est inférieure à la valeur Entrée et elle est éliminée si la probabilité est supérieure à la valeur Elimination. Pour remplacer les paramètres par défaut, indiquez des valeurs entières positives pour Entrée et Elimination. La valeur Entrée doit être inférieure à Elimination.

Césure du classement. Vous permet de définir la césure pour les observations de la classification. Les observations avec des prévisions qui excèdent la limite de classification sont classées positives tandis que celles dont les prévisions sont inférieures à la limite sont classées négatives. Pour modifier la valeur par défaut, entrez une valeur entre 0.01 et 0.99.

Maximum des itérations : Vous permet de modifier le nombre maximal d'itérations du modèle avant interruption.

Inclure terme constant dans le modèle. Vous permet d'indiquer si le modèle doit inclure un terme constant. Si cette option est désactivée, la constante est égale à 0.

Fonctions supplémentaires de la commande REGRESSION LOGISTIQUE

Le langage de syntaxe de commande vous permet aussi de :

- Identifier le résultat en fonction des observations par les valeurs ou les intitulés d'une variable.
- Commander l'espacement des rapports d'itération. Plutôt que d'imprimer les estimations après chaque itération, vous pouvez demander les estimations après chaque *énième* itération.
- Modifier les critères d'interruption d'une itération et de contrôle de la redondance.
- Spécifiez une liste de variables pour les listes par observations.
- Garder une trace en plaçant les données de chaque groupe de fichiers scindés dans un fichier vierge au cours du traitement.

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

Chapitre
3

Régression logistique multinomiale

La régression logistique multinomiale est utile dans le cas où vous souhaitez classer des objets en fonction des valeurs d'un groupe de variables prédites. Ce type de régression est similaire à la régression logistique, mais s'avère plus général puisque la variable dépendante n'est pas limitée à deux modalités.

Exemple : Afin de mieux rentabiliser la commercialisation de leurs films, les studios souhaitent prévoir le type de film que les cinéphiles sont susceptibles d'aller voir. En effectuant une régression logistique multinomiale, le studio peut déterminer l'impact de l'âge, du sexe et de la situation de famille d'une personne sur les types de films qu'elle préfère. Le studio peut alors orienter la campagne promotionnelle d'un film particulier en fonction du groupe de spectateurs susceptibles d'aller le voir.

Statistiques : Historique des itérations, coefficients de paramètre, covariance asymptotique et matrices de corrélation, tests du rapport de vraisemblance pour les effets de modèle et les effets partiels, –2 log-vraisemblance. Qualité d'ajustement du Khi-deux de Pearson et de déviance. R^2 de Cox et Snell, de Nagelkerke et de McFadden. Classification : effectifs observés par rapport aux effectifs prédits par modalité de réponse. Tableau croisé : effectifs observés et prédits (avec résidus) et proportions par paramètre des covariables et par modalité de réponse.

Méthodes. Un modèle logit multinomial est ajusté pour le modèle factoriel complet ou pour un modèle défini par l'utilisateur. L'estimation des paramètres est effectuée au moyen d'un algorithme itératif calculant le maximum de vraisemblance.

Données. La variable dépendante doit être qualitative. Les variables indépendantes peuvent correspondre à des facteurs ou à des covariables. En général, les facteurs doivent être des variables qualitatives et les covariables, des variables continues.

Hypothèses : On suppose que les odds ratios de deux modalités quelconques sont indépendants de toutes les autres modalités de réponse. Par exemple, lorsqu'un nouveau produit est introduit sur un marché, ce postulat signifie que les parts de marché de tous les autres produits sont toutes affectées proportionnellement de la même façon. En outre, d'après un paramètre de covariable, les réponses sont supposées correspondre à des variables multinomiales indépendantes.

Obtention d'une régression logistique multinomiale

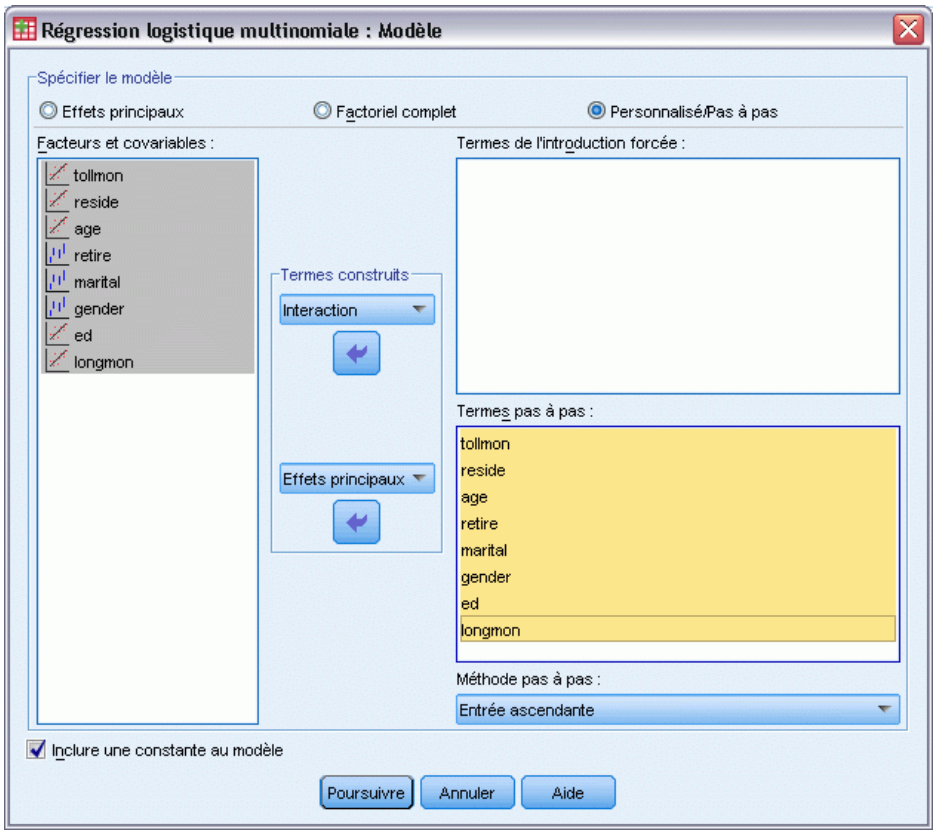
- A partir des menus, sélectionnez :
  Analyse > Régression > Logistique multinomiale...

Figure 3-1
Boîte de dialogue Régression logistique multinomiale

- Sélectionnez une variable dépendante.
- Les facteurs sont facultatifs et peuvent être numériques ou qualitatifs.
- Les covariables sont facultatives, mais doivent être numériques si elles sont spécifiées.

Régression logistique multinomiale

Figure 3-2
Boîte de dialogue Régression logistique multinomiale : Modèle



Par défaut, la procédure Régression logistique multinomiale crée un modèle contenant des effets principaux de covariable et de facteur, mais vous pouvez spécifier un modèle personnalisé ou choisir un modèle pas à pas dans cette boîte de dialogue.

Spécifier le modèle : Un modèle comportant des effets principaux contient des effets principaux de covariable et de facteur, mais aucun effet d'interaction. Un modèle factoriel complet contient tous les effets principaux et toutes les interactions inter-facteurs. Il ne contient pas de d'interactions de covariable. Vous pouvez créer un modèle personnalisé pour définir des sous-groupes d'interactions de facteurs ou de covariables, ou demander une sélection pas à pas de termes de modèle.

Facteurs et covariables : §FM Les facteurs et les covariables sont répertoriés.

Termes de l'introduction forcée : Les termes ajoutés à la liste d'introduction forcée sont systématiquement inclus dans le modèle.

Termes pas à pas : Les termes ajoutés à la liste pas à pas sont inclus dans le modèle, en fonction de l'une des méthodes pas à pas suivantes sélectionnées par l'utilisateur :

- **Entrée ascendante :** A la première étape de cette méthode, le modèle ne contient aucun terme pas à pas. A chaque étape, le terme le plus significatif est ajouté au modèle jusqu'à ce qu'aucun terme pas à pas exclu du modèle n'ait de contribution statistiquement significative s'il est inséré dans ce modèle.
- **Elimination descendante :** La première étape de cette méthode consiste à insérer dans le modèle tous les termes de la liste pas à pas. A chaque étape, le terme pas à pas le moins significatif est supprimé du modèle jusqu'à ce que tous les termes pas à pas restants aient une contribution statistiquement significative pour ce modèle.
- **Pas à pas ascendant.** La première étape de cette méthode consiste à sélectionner le modèle par la méthode d'introduction ascendante. A partir de là, l'algorithme alterne entre élimination descendante des termes pas à pas du modèle et introduction ascendante des termes exclus de ce modèle. Ce processus se poursuit jusqu'à ce que plus aucun terme ne réponde aux critères d'ajout ou de suppression.
- **Pas à pas descendante** : La première étape de cette méthode consiste à sélectionner le modèle par la méthode d'élimination descendante. A partir de là, l'algorithme alterne entre introduction ascendante des termes exclus du modèle et élimination descendante des termes pas à pas de ce modèle. Ce processus se poursuit jusqu'à ce que plus aucun terme ne réponde aux critères d'ajout ou de suppression.

Inclure ordonnée à l'origine dans le modèle : Cette option vous permet d'inclure ou d'exclure une constante pour le modèle.

Termes construits

Pour les facteurs et covariables sélectionnés :

Interaction : Crée le terme d'interaction du plus haut niveau de toutes les variables sélectionnées.

Effets principaux : Crée un terme d'effet principal pour chaque variable sélectionnée.

Toutes d'ordre 2 : Crée toutes les interactions d'ordre 2 possibles des variables sélectionnées.

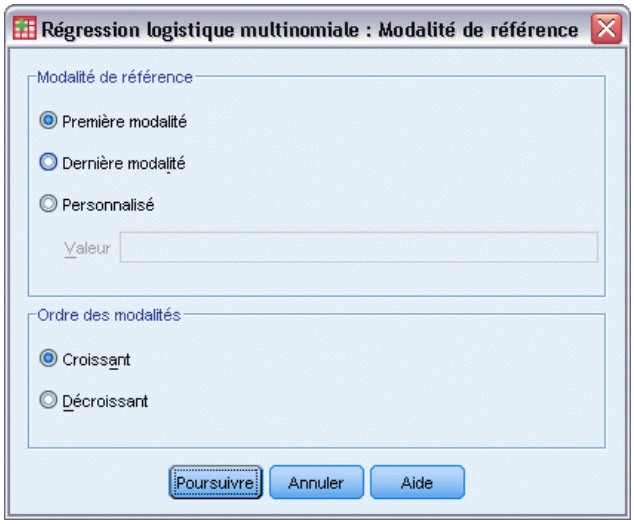
Toutes d'ordre 3 : Crée toutes les interactions d'ordre 3 possibles des variables sélectionnées.

Toutes d'ordre 4 : Crée toutes les interactions d'ordre 4 possibles des variables sélectionnées.

Toutes d'ordre 5 : Crée toutes les interactions d'ordre 5 possibles des variables sélectionnées.

Régression logistique multinomiale : Modalité de référence

Figure 3-3
Boîte de dialogue Régression logistique multinomiale : Modalité de référence



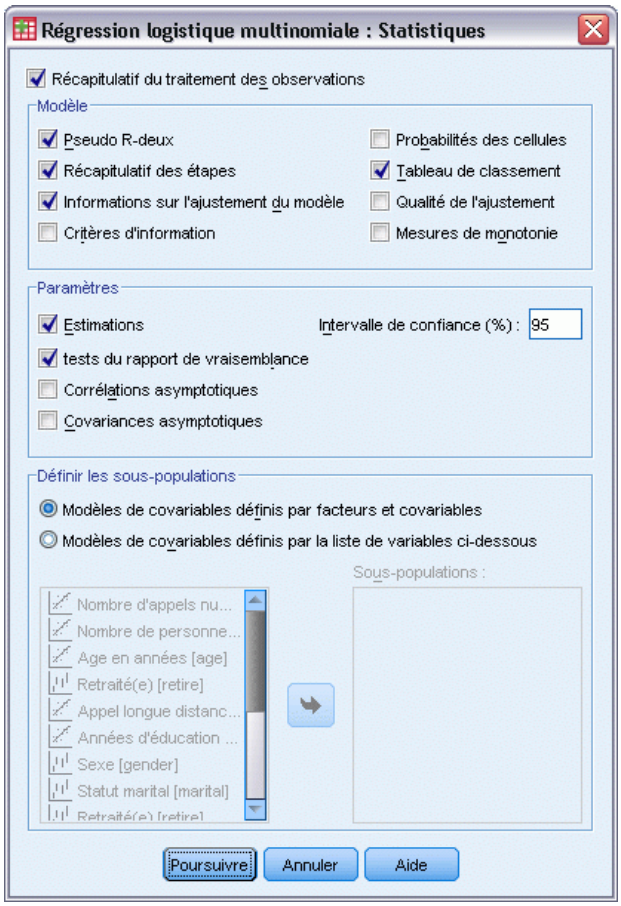
Par défaut, la procédure Régression logistique multinomiale utilise la dernière modalité comme modalité de référence. Cette boîte de dialogue vous permet de contrôler la modalité de référence et le type de tri des modalités.

Modalité de référence : Spécifiez la première ou la dernière modalité, ou une modalité personnalisée.

Ordre des modalités : Dans l'ordre croissant, la valeur minimale définit la première modalité et la valeur maximale, la dernière modalité. Dans l'ordre décroissant, la valeur maximale définit la première modalité et la valeur minimale, la dernière modalité.

Régression logistique multinomiale : Statistiques

Figure 3-4
Boîte de dialogue Régression logistique multinomiale : Statistiques



Les statistiques pouvant être définies pour la régression logistique multinomiale sont les suivantes :

Récapitulatif du traitement des observations : Ce tableau contient les informations relatives aux variables qualitatives fournies.

Modèle : Statistiques du modèle global.

- **Pseudo R-deux.** Imprime les statistiques R^2 de Cox et Snell, de Nagelkerke et de McFadden.
- **Récapitulatif des étapes :** Ce tableau récapitule les effets ajoutés à chaque étape d'une méthode pas à pas ou supprimés de cette dernière. Il n'est créé que si un modèle pas à pas est spécifié dans la boîte de dialogue Modèle
- **Informations sur l'ajustement du modèle :** Ce tableau compare les modèles ajustés et les modèles avec constante seulement ou les modèles nuls.
- **Critères d'information.** Ce tableau imprime le critère d'information d'Akaike (AIC) et le critère d'information bayésien de Schwarz (BIC).
- **Probabilités des cellules :** Imprime un tableau des effectifs observés et des effectifs théoriques (avec résidu), et des proportions par paramètre de covariable et par modalité de réponse.

- **Tableau de classement :** Imprime un tableau comparatif des réponses observées et des réponses prédites.
- **Qualité d'ajustement de statistiques de Khi-deux :** Imprime les statistiques Khi-deux de Pearson et Khi-deux du rapport de vraisemblance. Les statistiques sont calculées pour les paramètres de covariable déterminés par tous les facteurs et covariables ou par un sous-ensemble de facteurs et de covariables défini par l'utilisateur.
- **Mesures de monotonicité.** Affiche un tableau contenant des informations sur les nombres de paires concordantes, de paires discordantes et de paires liées. Le D de Somers, le Gamma de Goodman et Kruskal, le Tau-a de Kendall et l'Indice de concordance C apparaissent également dans ce tableau.

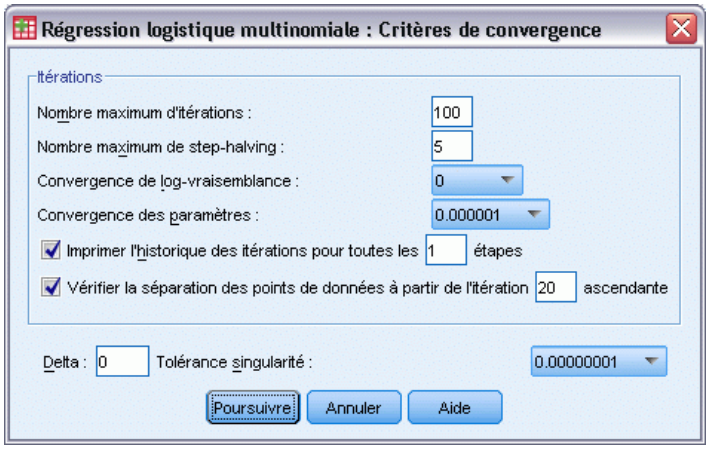
Paramètres : Statistiques liées aux paramètres du modèle.

- **Estimations :** Imprime les estimations des paramètres du modèle, avec un niveau de confiance défini par l'utilisateur.
- **Test du ratio de vraisemblance :** Imprime les tests du rapport de vraisemblance pour les effets partiels du modèle. Le test du modèle global est imprimé automatiquement.
- **Corrélations asymptotiques :** Imprime la matrice de corrélation des estimations de paramètres.
- **Covariances asymptotiques :** Imprime la matrice de covariance des estimations de paramètres.

Définir les sous-populations : Cette option vous permet de sélectionner un sous-ensemble de facteurs et de covariables afin de définir les paramètres de covariable utilisés par les probabilités des cellules et par les tests de qualité d'ajustement.

Régression logistique multinomiale : Critères

Figure 3-5
Boîte de dialogue Régression logistique multinomiale : Critères de convergence



Les critères pouvant être définis pour la régression logistique multinomiale sont les suivants :

Itérations : Cette option vous permet d'indiquer le nombre de fois où vous souhaitez répéter l'algorithme, le nombre maximal d'étapes de la procédure de méthode dichotomique, les tolérances de convergence relatives aux modifications de la log-vraisemblance et des paramètres, l'effectif

des impressions de l'état d'avancement de l'algorithme itératif, ainsi que l'itération à laquelle la procédure doit commencer à rechercher une séparation complète ou quasi-complète des données.

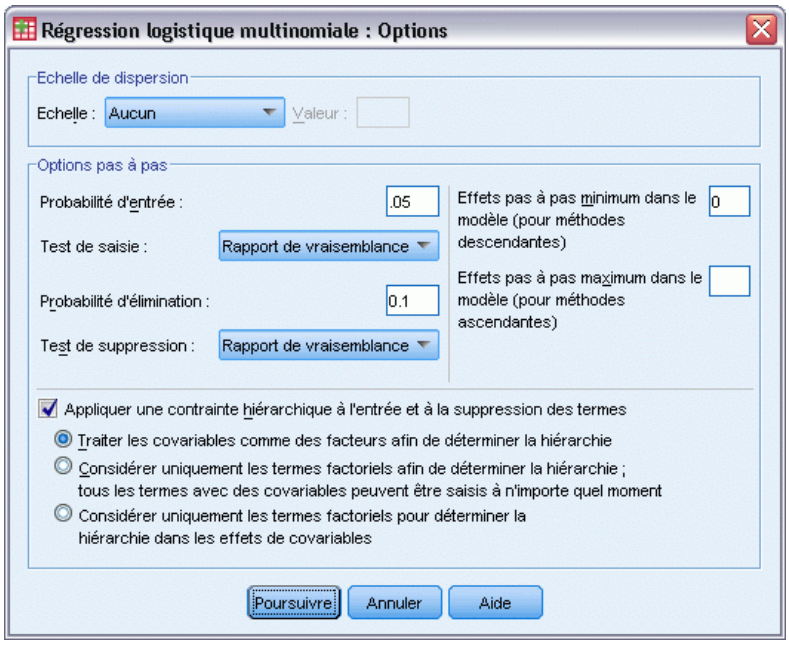
- **Convergence de log-vraisemblance :** La convergence est supposée si la variation absolue de la fonction log-vraisemblance est inférieure à une valeur donnée. Le critère n'est pas utilisé si la valeur est 0. Indiquez une valeur non négative.
- **Convergence des paramètres.** La convergence est prise en compte si la modification absolue des estimations du paramètre est inférieure à cette valeur. Le critère n'est pas utilisé si la valeur est 0.

Delta. Cette option vous permet de définir une valeur non négative inférieure à 1. Cette valeur est ajoutée à chacune des cellules vides du tableau croisé des modalités de réponse par paramètre de covariable. Ceci vous permet de stabiliser l'algorithme et d'éviter les estimations biaisées.

Tolérance singularité : Cette option vous permet de définir la tolérance utilisée lors du contrôle des particularités.

Options de régression logistique multinomiale

Figure 3-6
Boîte de dialogue Régression logistique multinomiale : Options



Les statistiques pouvant être définies pour la régression logistique multinomiale sont les suivantes :

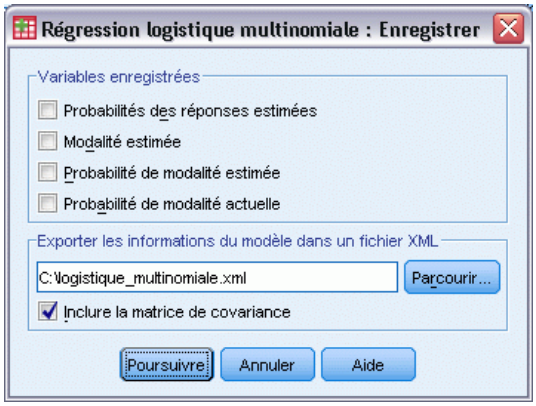
Echelle de dispersion : Cette option vous permet de définir la valeur d'échelle de dispersion qui sera utilisée pour corriger l'estimation de la matrice de covariance des paramètres. L'option Déviance estime la valeur d'échelle au moyen de la statistique fonction de déviance (Khi-deux du rapport de vraisemblance). L'option Pearson estime la valeur d'échelle à l'aide de la statistique Khi-deux de Pearson. Vous pouvez également spécifier votre propre valeur d'échelle. Il doit s'agir d'une valeur numérique positive.

Options pas à pas : Ces options vous permettent de contrôler les critères statistiques lorsque des méthodes pas à pas servent à créer un modèle.Ils sont ignorés sauf si un modèle pas à pas est spécifié dans la boîte de dialogue Modèle

- **Probabilité d'entrée :** Il s'agit de la probabilité de la statistique du rapport de vraisemblance pour l'entrée de variables. La facilité avec laquelle une variable est ajoutée au modèle dépend directement de la valeur de la probabilité fournie. Plus cette valeur est élevée, plus la variable a de chances d'être insérée dans le modèle. Ce critère est ignoré sauf si la méthode d'introduction ascendante, ou la méthode pas à pas ascendante ou descendante est sélectionnée.
- **Test de saisie.** Il s'agit de la méthode permettant de saisir des termes selon des méthodes détaillées étape par étape. Choisissez entre le test du rapport de vraisemblance et le test de significativité. Ce critère est ignoré sauf si la méthode d'introduction ascendante, ou la méthode pas à pas ascendante ou descendante est sélectionnée.
- **Probabilité d'élimination :** Il s'agit de la probabilité de la statistique du rapport de vraisemblance pour la suppression de variables. La facilité avec laquelle une variable est conservée dans le modèle dépend directement de la valeur de la probabilité fournie. Plus cette valeur est élevée, plus la variable a de chances de rester dans le modèle. Ce critère est ignoré sauf si la méthode d'élimination descendante, ou la méthode pas à pas ascendante ou descendante est sélectionnée.
- **Test de suppression.** Il s'agit de la méthode permettant de supprimer des termes selon des méthodes détaillées étape par étape. Choisissez entre le test du rapport de vraisemblance et le test de Wald. Ce critère est ignoré sauf si la méthode d'élimination descendante, ou la méthode pas à pas ascendante ou descendante est sélectionnée.
- **Effets pas à pas minimum dans le modèle.** Lorsque la méthode pas à pas descendante ou la méthode d'élimination descendante est utilisée, cette option spécifie le nombre minimal de termes à inclure dans le modèle. La constante n'est pas considérée comme terme de modèle.
- **Effets pas à pas maximum dans le modèle.** Lorsque la méthode pas à pas ascendante ou la méthode d'introduction ascendante est utilisée, cette option spécifie le nombre maximal de termes à inclure dans le modèle. La constante n'est pas considérée comme terme de modèle.
- **Appliquer une contrainte hiérarchique à l'entrée et à l'élimination des termes :** Cette option vous permet d'indiquer si des restrictions doivent s'appliquer à l'ajout de termes de modèle. La hiérarchie exige que, pour tout terme à inclure, l'ensemble des termes de niveau inférieur appartenant à ce terme figure avant tout dans le modèle. Par exemple, si cette exigence de la hiérarchie est appliquée, les facteurs *Situation familiale* et *Sexe* doivent être contenus dans le modèle pour que l'interaction *Situation familiale*Sexe* puisse être ajoutée. Les trois boutons radio déterminent le rôle que jouent les covariables dans l'établissement de la hiérarchie.

Régression logistique multinomiale : Enregistrer

Figure 3-7
Boîte de dialogue Régression logistique multinomiale : Enregistrer



La boîte de dialogue Enregistrer vous permet d'enregistrer des variables dans le fichier de travail et d'exporter les informations du modèle vers un fichier externe.

Variables enregistrées :

- **Probabilités des réponses estimées :** Il s'agit des probabilités estimées de classement d'un type de facteur/covariable dans les modalités de réponse. Il y a autant de probabilités estimées que de modalités de variables de réponse ; jusqu'à 25 probabilités seront enregistrées.
- **Modalité estimée :** Il s'agit de la modalité de réponse dont le nombre de probabilités théorique est le plus élevé pour un type de facteur/covariable.
- **Probabilité de modalité estimée :** Il s'agit du nombre maximum de probabilités de réponses estimées.
- **Probabilité de modalité actuelle :** Il s'agit de la probabilité estimée sur un modèle de classement d'un type de facteur/covariable dans la modalité observée.

Exporter les informations du modèle dans un fichier XML : Les estimations de paramètres et leurs covariances (facultatif) sont exportées vers le fichier spécifié au format XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.

Fonctionnalités supplémentaires de la commande NOMREG

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier la modalité de référence de la variable dépendante.
- d'inclure les observations avec valeurs manquantes spécifiées par l'utilisateur ;
- personnaliser les tests d'hypothèse en spécifiant des hypothèses nulles comme combinaisons linéaires de paramètres.

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

Chapitre
4

Modèles de choix binaire

Cette procédure mesure la relation entre l'intensité d'un stimulus et la proportion des observations montrant une certaine réponse au stimulus. Elle est utile lorsque vous avez un résultat dichotomique qu'on pense être influencé ou causé par des niveaux de certaines variables indépendantes. Elle est de ce fait bien adaptée aux données expérimentales. Cette procédure vous permet d'estimer la force d'un stimulus requise pour induire une certaine proportion de réponses, telle que la dose moyenne efficace.

Exemple : Quelle est l'efficacité d'un nouveau pesticide contre les fourmis et quelle concentration doit-on utiliser ? Vous devez mener une expérience dans laquelle vous exposez des échantillons de fourmis à différentes concentrations de pesticide et vous enregistrez le nombre de fourmis tuées et le nombre de fourmis exposées. En appliquant l'analyse des modèles de choix binaire à ces données, vous pouvez déterminer la force de la relation entre la concentration et la destruction de fourmis et la proportion de pesticide nécessaire si vous souhaitez être sûr de vous débarrasser de, disons, 95 % des fourmis exposées.

Statistiques : Coefficients de régression et erreurs standard, constante et erreur standard, Khi-deux de la qualité de l'ajustement de Pearson, fréquences attendues et théoriques et intervalle de confiance pour les niveaux efficients des variables indépendantes. Diagrammes : diagrammes de réponse transformés.

Cette procédure utilise les algorithmes proposés et mis en œuvre dans NPSOL® par Gill, Murray, Saunders & Wright pour estimer les paramètres du modèle.

Données : Pour chaque valeur de la variable indépendante (ou chaque combinaison de valeurs de plusieurs variables indépendantes), votre variable de réponse doit être l'effectif du nombre d'observations avec celles des valeurs qui montrent la réponse d'intérêt, et la variable observée totale doit être un effectif du nombre total d'observations avec celles des valeurs de la variable indépendante. La variable active doit être qualitative, codée sous la forme de nombres entiers.

Hypothèses : Les observations doivent être indépendantes. Si vous avez un grand nombre de valeurs pour les variables indépendantes relatives au nombre d'observations, comme c'est peut-être le cas dans votre étude, le Khi-deux et les statistiques de la qualité de l'ajustement ne sont peut-être pas valables.

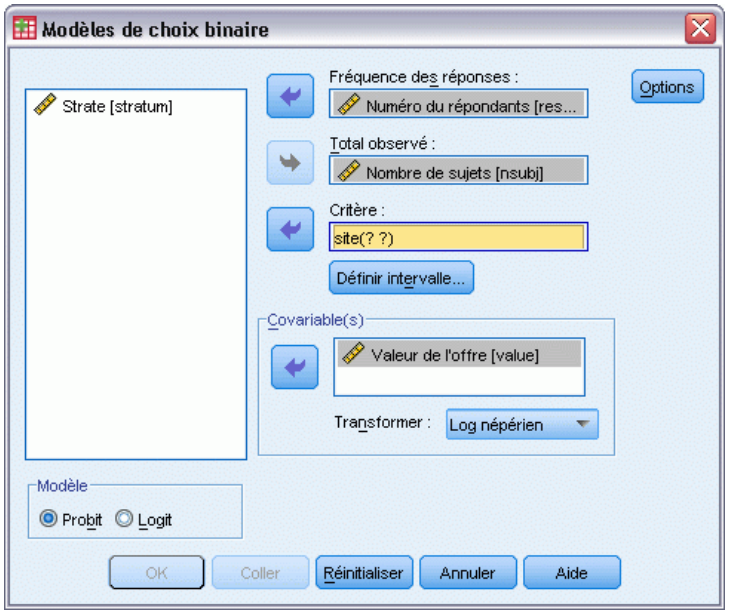
Procédures apparentées : Les modèles de choix binaire sont étroitement liés à la régression logistique. En fait, si vous sélectionnez la transformation logit, cette procédure calculera essentiellement une régression logistique. En général, les modèles de choix binaire s'adaptent à des plans d'expériences, tandis que la régression logistique est plus appropriée pour des études par observation. Les différences au niveau du résultat reflètent ces différentes emphases. La procédure des modèles de choix binaire offre des estimations de valeurs effectives pour divers niveaux de

réponse (incluant la dose effective médiane), tandis que la procédure de la régression logistique offre des estimations des odds ratios pour les variables indépendantes.

Obtenir des modèles de choix binaire

- A partir des menus, sélectionnez :
  Analyse > Régression > Modèles de choix binaire...

Figure 4-1
Boîte de dialogue Modèles de choix binaire

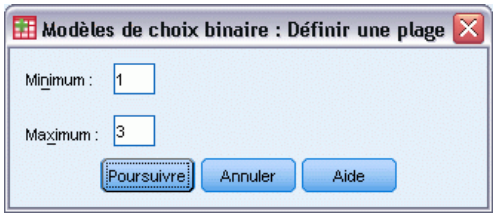


- Sélectionnez une variable de fréquence des réponses. Cette variable indique le nombre d'observations présentant une réponse au stimulus test. Les valeurs de cette variable ne peuvent pas être négatives.
- Sélectionnez une variable totale observée. Cette variable indique le nombre d'observations auxquelles le stimulus a été appliqué. Les valeurs de cette variable ne peuvent pas être négatives et ne peuvent pas être inférieures à la variable de fréquence de réponse pour chaque observation.

  Vous pouvez également sélectionnez une variable active. Sinon, cliquez sur Définir intervalle pour définir les groupes.
- Sélectionnez une ou plusieurs covariables. Cette variable contient le niveau du stimulus appliqué à chaque observation. Si vous souhaitez transformer la covariable, sélectionnez une transformation à partir de la liste déroulante Transformation. Si vous n'appliquez aucune transformation et qu'il existe un groupe de contrôle, ce groupe de contrôle est alors inclus dans l'analyse.
- Sélectionnez le modèle Probit ou le modèle Logit.
  - **Modèle probit.** Applique la transformation probit (inverse de la fonction de distribution normale standard cumulée) aux proportions de réponses.
  - **Modèle logit.** Applique la transformation logit (probabilités logarithmiques) aux proportions de réponses.

Modèles de choix binaire : définir un intervalle

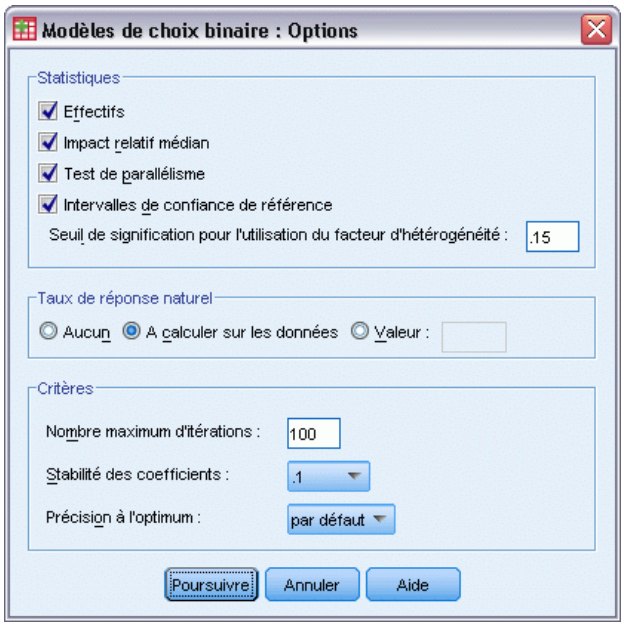
Figure 4-2
Boîte de dialogue Modèles de choix binaire : Définir intervalle



Cela vous permet de spécifier les variables actives qui seront analysées. Les niveaux de facteur doivent être codés sous la forme de nombres entiers consécutifs, et tous les niveaux que vous indiquez doivent être analysés.

Options des modèles de choix binaire

Figure 4-3
Boîte de dialogue Modèles de choix binaire : Options



Vous pouvez spécifier certaines options pour vos modèles de choix binaire :

Statistiques : Vous permet de demander les options statistiques suivantes : Fréquences, Impact relatif médian, Test de parallélisme, Intervalles de confiance de référence.

- **Impact relatif médian.** Affiche le ratio des impacts moyens pour chaque paire de niveaux de facteurs. Montre également les intervalles de confiance à 95 % pour chacun des impacts relatifs médians. Les impacts relatifs médians ne sont pas disponibles s'il n'y a pas de variable active ou s'il existe plusieurs covariables.

- **Test de parallélisme.** Test de l'hypothèse selon laquelle tous les niveaux de facteur ont une pente commune.
- **Intervalles de confiance de référence.** Intervalles de confiance pour que le dosage de l'agent nécessaire produise une certaine probabilité de réponse.

Les intervalles de confiance de référence et l'impact relatif médian ne sont pas disponibles si vous avez sélectionné plus d'une covariable. L'impact relatif médian et le test de parallélisme sont disponibles uniquement lorsque vous avez sélectionné une variable active.

Taux de réponses spontanées. Vous permet d'indiquer un taux de réponse spontané même en l'absence de stimulus. Les options disponibles sont Aucun, Calculer à partir des données ou Valeur.

- **Calculer à partir des données.** Calcule le taux de réponse naturel à partir des données de l'échantillon. Vos données doivent contenir une observation représentant le niveau de contrôle pour lequel la valeur des covariables est 0. Probit estime le taux de réponse naturel à partir de la proportion de réponses pour le niveau de contrôle comme une valeur initiale.
- **Valeur.** Définit le taux de réponse naturel du modèle (sélectionnez cette option lorsque vous souhaitez connaître le taux de réponse naturel à l'avance). Tapez la proportion de réponse naturelle (cette proportion doit être inférieure à 1). Si, par exemple, une réponse existe dans 10 % des cas lorsque le stimulus est de 0, tapez 0,10.

Critères : Vous permet de commander les paramètres de l'algorithme itératif d'estimations. Vous pouvez passer outre les valeurs par défaut pour le maximum des itérations, la stabilité des coefficients et la précision à l'optimum.

Fonctions supplémentaires de la commande NLR

Le langage de syntaxe de commande vous permet aussi de :

- Demander une analyse parmi les deux analyses, Probit et Logit.
- Commander le traitement des valeurs manquantes.
- Transformer les covariables par des bases différentes de la base 10 ou des logarithmes naturels.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
5

Régression non linéaire

La régression non linéaire est une méthode permettant de déterminer un modèle non linéaire de relation entre la variable dépendante et un groupe de variables indépendantes. A l'inverse de la régression linéaire classique, qui se limite aux modèles linéaires de prévision, la régression non linéaire peut élaborer des modèles avec des relations arbitraires entre variables dépendantes et indépendantes. Elle emploie pour cela des algorithmes itératifs d'estimation. Remarquez que cette procédure n'est pas indispensable pour les simples modèles polynomiaux de forme : Y = A + BX**2. Si on pose W = X**2, il s'agit d'un simple modèle linéaire de type Y = A + BW qui peut être estimé à partir de méthodes traditionnelles comme la procédure de régression linéaire.

Exemple : Peut-on estimer l'évolution de la population par rapport au temps ? Un diagramme de dispersion montre qu'il semble y avoir une forte relation entre la population et le temps mais cette relation n'est pas linéaire. Il faut donc employer des méthodes de prévision particulières de la procédure de régression non linéaire. En déterminant une équation appropriée, tel qu'un modèle logistique d'évolution de la population, nous pouvons obtenir une bonne approximation du modèle, ce qui nous permet de prévoir la population à des dates pour lesquelles elle n'a pas encore été mesurée.

Statistiques : Pour chaque itération : estimations des paramètres et somme résiduelle des carrés. Pour chaque modèle : somme des carrés pour la régression, le résidu, le total correct ou incorrect, les estimations, les erreurs standard asymptotiques et la matrice de corrélation asymptotique des estimations.

Remarque : La régression non linéaire restreinte utilise les algorithmes proposés et mis en oeuvre dans NPSOL® par Gill, Murray, Saunders et Wright pour estimer les paramètres du modèle.

Données : Les variables dépendantes et indépendantes doivent être quantitatives. Les variables qualitatives, comme la religion, la qualification, la zone de résidence, doivent être enregistrées sous forme de variables binaires (factices) ou sous de tout autre type de variables de contraste.

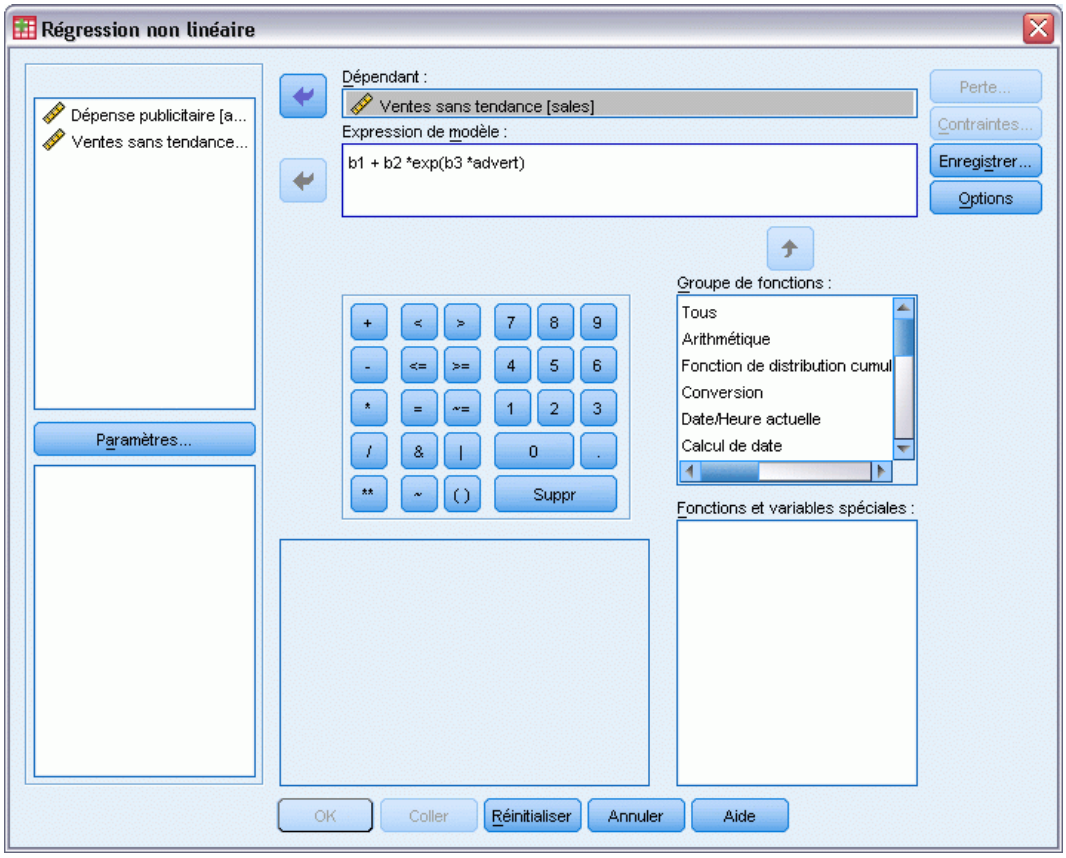
Hypothèses : Les résultats ne sont valides que si vous avez indiqué une fonction qui décrit correctement la relation entre les variables dépendantes et indépendantes. De surcroît, le choix des bonnes valeurs de départ est très important. Même si vous avez spécifié la forme fonctionnelle correcte du modèle, si vous utilisez de mauvaises valeurs de départ, votre modèle risque de ne pas réussir à converger et vous n'obtiendrez qu'un modèle optimal locale et non pas globale.

Procédures apparentées : De nombreux modèles qui n'apparaissent pas linéaires à première vue peuvent être transformés en modèles linéaires et analysés à l'aide une procédure de régression linéaire. Si vous n'êtes pas sûr du modèle à employer, la procédure d'ajustement de fonctions peut vous permettre d'identifier les relations fonctionnelles utiles dans vos données.

Obtenir une analyse de la régression non linéaire

▶ A partir des menus, sélectionnez :
Analyse > Régression > Non linéaire...

Figure 5-1
Boîte de dialogue Régression non linéaire



▶ Sélectionnez une variable dépendante (numérique) dans la liste des variables de votre ensemble de données actif.

▶ Pour définir l'expression du modèle, entrez l'expression dans le champ Modèle ou collez les composants (variables, paramètres, fonctions) dans le champ.

▶ Pour identifier les paramètres de votre modèle, cliquez sur Paramètres.

Un modèle segmenté (qui prend différentes formes dans les différentes parties du domaine) peut être spécifié à l'aide d'une logique conditionnelle au sein d'une même déclaration de modèle.

Logique conditionnelle (régression non linéaire)

Vous pouvez spécifier un modèle segmenté à l'aide d'une logique conditionnelle. Pour employer une logique conditionnelle dans l'expression d'un modèle ou une fonction de perte, vous formez la somme d'une série de termes pour chaque condition. Chaque terme contient une expression

logique (entre parenthèses) multipliée par l'expression qui doit résulter lorsque l'expression logique est vraie.

Par exemple, considérez un modèle segmenté qui est égal à 0 pour X<= 0, à X pour 0<X<1 et à 1 pour X>=1. L'expression de ce modèle est :

(X<=0)*0 + (X>0 & X < 1)*X + (X>=1)*1.

Les expressions logiques entre parenthèses ont toutes pour résultat 1 (vrai) ou 0 (faux). Donc :

Si X<= 0, l'expression se réduit à 1*0 + 0*X + 0*1 = 0.

Si 0<X<1, elle se réduit à 0*0 + 1*X +0*1 = X.

Si X>=1, elle devient 0*0 + 0*X + 1*1 = 1.

Des exemples plus complexes peuvent se construire facilement en substituant les différentes expressions logiques et les expressions de sortie. Gardez en mémoire que les doubles inégalités, telles que 0<X<1, doivent être écrites sous forme d'expressions composées de type (X>0 & X < 1).

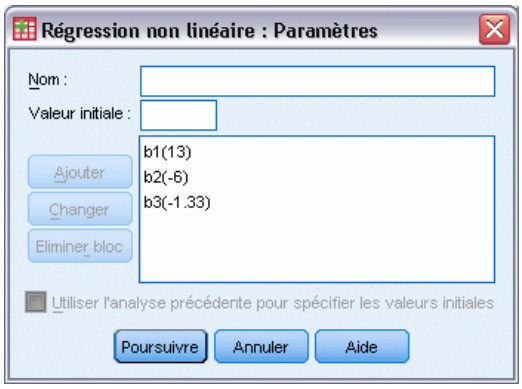
Les variables chaîne peuvent être utilisées dans les expressions logiques :

(ville='Paris')*pouvach + (ville='Maubeuge')*0.59*pouvach

Cela produit l'expression (la valeur de la variable *pouvach*) pour les Parisiens et une autre (59 % de cette valeur) pour les habitants de Maubeuge. Les constantes alphanumériques doivent être présentées entre guillemets ou apostrophes, comme dans cet exemple.

Paramètres de régression non linéaire

Figure 5-2
Boîte de dialogue Régression non linéaire : Paramètres



Les paramètres constituent les parties de votre modèle que la procédure de régression non linéaire estime. Ces paramètres peuvent être des constantes additives, des coefficients multiplicateurs, des exposants ou des valeurs utilisés dans les fonctions d'évaluation. Tous les paramètres que vous avez définis apparaissent (avec leurs valeurs initiales) dans la liste Paramètres de la boîte de dialogue.

Nom. Vous devez attribuer un nom à chaque paramètre. Ce nom doit être un nom de variable valide et doit être le nom utilisé dans l'expression de modèle de la boîte de dialogue principale.

Valeur initiale. Vous permet de spécifier une valeur initiale pour le paramètre, de préférence aussi proche que possible de la solution finale escomptée. De mauvaises valeurs initiales peuvent entraîner des problèmes de convergence (impossibilité de converger ou convergence locale plutôt que globale).

Utiliser l'analyse précédente pour spécifier les valeurs initiales. Si vous avez déjà exécuté une régression non linéaire à partir de cette boîte de dialogue, vous pouvez sélectionner cette option pour obtenir les valeurs initiales des paramètres à partir des valeurs de la précédente exécution. Cela vous permet de continuer la recherche lorsque l'algorithme converge lentement. (Les valeurs initiales de départ apparaissent toujours dans la liste Paramètres de la boîte de dialogue principale.)

Remarque : Cette sélection persiste dans la boîte de dialogue pour le reste de la session. Si vous changez de modèle, assurez-vous de le désélectionner.

Modèles courants de régression non linéaire

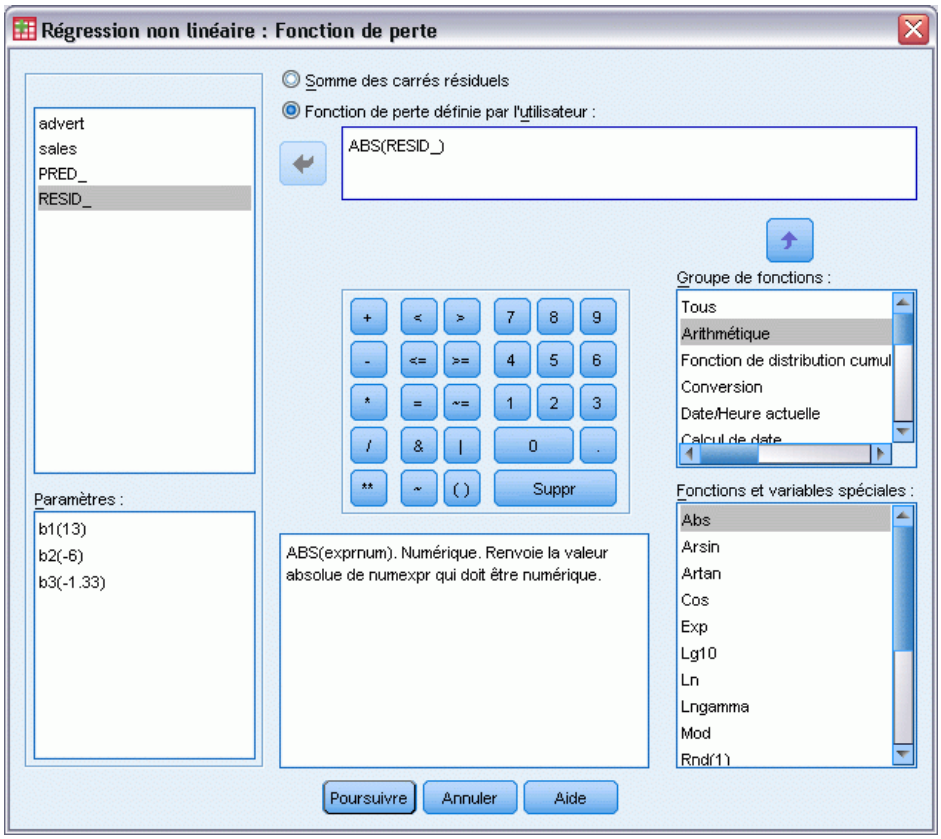
Le tableau suivant présente un exemple de syntaxe de plusieurs modèles de régression non linéaire. Un modèle choisi au hasard a peu de chance de s'adapter à vos données. Les valeurs de départ appropriées pour les paramètres sont indispensables et certains modèles requièrent des contraintes afin de converger.

Table 5-1
Exemple de syntaxe

Nom	Expression
Régression asymptotique	b1 + b2 *exp(b3 * x)
Régression asymptotique	b1 –(b2 *(b3 ** x))
Densité	(b1 + b2 * x)**(–1/ b3)
Gauss	b1 *(1– b3 *exp(–b2 * x **2))
Gompertz	b1 *exp(–b2 * exp(–b3 * x))
Johnson-Schumacher	b1 *exp(–b2 / (x + b3))
Log modifié	(b1 + b3 * x) ** b2
Log logistique	b1 – ln(1+ b2 *exp(–b3 * x))
Loi des réponses décroissantes de Metcherlich	b1 + b2 *exp(–b3 * x)
Michaelis Menten	b1* x /(x + b2)
Morgan-Mercer-Florin	(b1 * b2 + b3 * x ** b4)/(b2 + x ** b4)
Peal-Reed	b1 /(1+ b2 *exp(–(b3 * x + b4 * x **2 + b5 * x **3)))
Ratio cubique	(b1 + b2 * x + b3 * x **2 + b4 * x **3)/(b5 * x **3)
Ratio quadratique	(b1 + b2 * x + b3 * x **2)/(b4 * x **2)
Richards	b1 /((1+ b3 *exp(–b2 * x))**(1/ b4))
Verhulst	b1 /(1 + b3 * exp(– b2 * x))
Von Bertalanffy	(b1 ** (1 – b4) – b2 * exp(–b3 * x)) ** (1/(1 –b4))
Weibull	b1 – b2 *exp(– b3 * x ** b4)
Densité de rendement	(b1 + b2 * x + b3 * x **2)**(–1)

Fonction de perte de la régression non linéaire

Figure 5-3
Boîte de dialogue Régression non linéaire : Fonction de perte



La **fonction de perte** dans la régression non linéaire est la fonction minimisée par l'algorithme. Sélectionnez soit Somme des carrés des résidus pour minimiser la somme des carrés résiduels, soit Fonction de perte spécifiée par l'utilisateur pour minimiser une fonction différente.

Si vous sélectionnez Fonction de perte spécifiée par l'utilisateur, vous devez définir la fonction de perte dont la somme (sur toutes les observations) doit être minimisée par le choix des valeurs du paramètre.

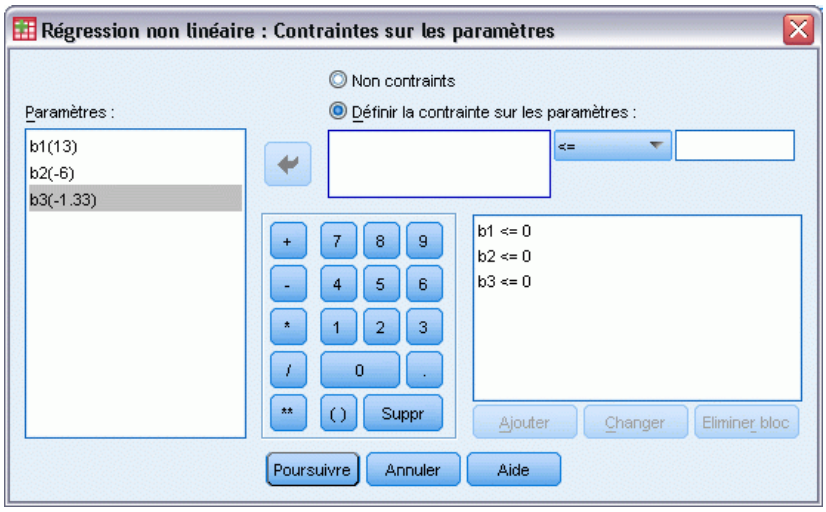
- La plupart des fonctions de perte impliquent la variable spéciale *RESID_*, qui représente le résidu. (La fonction de perte par défaut Somme des carrés résiduels doit être saisie explicitement sous la forme RESID_**2.) Si vous avez besoin d'employer la valeur prévisionnelle dans votre fonction de perte, cette valeur est égale à la variable dépendante moins le résidu.
- Il est possible de spécifier une fonction de perte conditionnelle à l'aide de la logique conditionnelle.

Vous pouvez soit taper une expression dans le champ de la fonction de perte personnalisée (spécifiée par l'utilisateur), soit coller les composants de cette expression dans le champ. Les constantes alphanumériques doivent être saisies entre guillemets ou apostrophes, tandis que les

constantes numériques doivent être en format Américain avec un point en tant que séparateur décimal.

Options de contraintes de la régression non linéaire

Figure 5-4
Boîte de dialogue Régression non linéaire Contraintes sur les paramètres



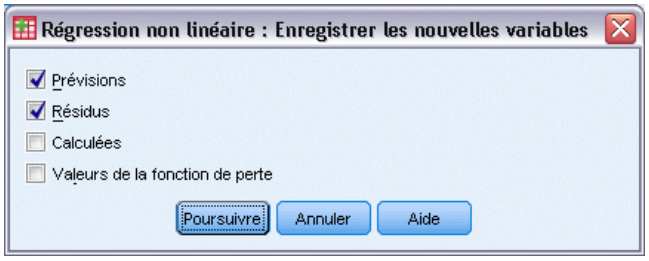
Une **contrainte** est une restriction émise sur les valeurs permises d'un paramètre au cours du processus itératif de recherche d'une solution. Les expressions linéaires sont évaluées avant chaque pas. Vous pouvez donc utiliser les contraintes linéaires pour éviter les pas qui risquent d'entraîner des dépassements positifs. Les expressions non linéaires sont évaluées après chaque pas.

Chaque équation ou inégalité requièrent les éléments suivants :

- Une expression impliquant au moins un paramètre dans le modèle. Saisissez l'expression ou employez le clavier qui vous permet de coller des nombres, des opérateurs ou des parenthèses dans une expression. Vous pouvez soit taper les paramètres requis avec le reste de l'expression ou les coller depuis la liste Paramètres sur la gauche. Vous ne pouvez pas utiliser de variables courantes dans une contrainte.
- Un des trois opérateurs logiques <=, = ou >=.
- Une constante numérique à laquelle l'expression est comparée à l'aide de l'opérateur logique. Tapez la constante. Les constantes numériques doivent être saisies en format Américain avec un point en tant que séparateur décimal.

Régression non linéaire : enregistrer les nouvelles variables

Figure 5-5
Boîte de dialogue Régression non linéaire : Enregistrer les nouvelles variables

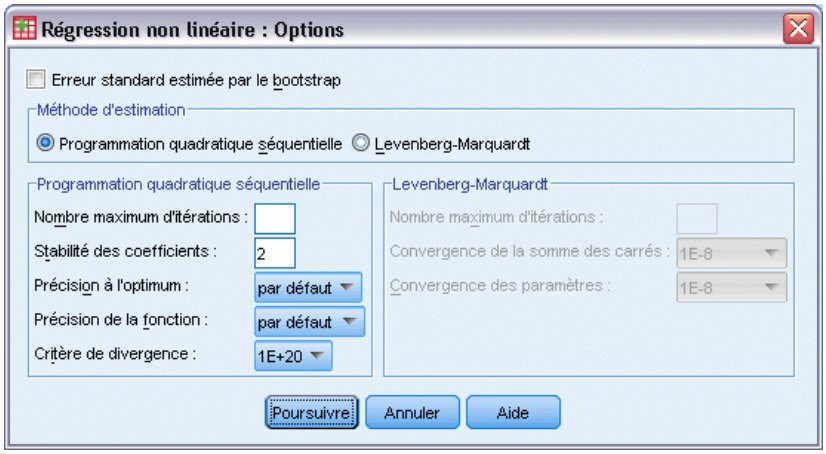


Vous pouvez enregistrer un certain nombre de nouvelles variables dans votre fichier de données actif. Les options disponibles sont Prévisions, Résidus, Calculées, et Valeurs de la fonction de perte. Ces variables peuvent servir dans les analyses suivantes pour tester l'adéquation du modèle ou pour identifier les observations problématiques.

- **Résidus.** Enregistre les résidus avec les noms de variable resid.
- **Prévisions.** Enregistre les valeurs estimées avec les noms de variable pred_.
- **Dérivées.** Une dérivée est enregistrée pour chacun des paramètres du modèle. Le nom d'une dérivée est formé à partir du préfixe d suivi des six premiers caractères du nom du paramètre.
- **Valeurs de la fonction de perte.** Cette option est accessible si vous spécifiez votre propre fonction de perte. Le nom de variable loss_ est affecté aux valeurs de la fonction de perte.

Options de régression non linéaire

Figure 5-6
Boîte de dialogue Régression non linéaire : Options



Ces options permettent de commander les différents aspects de votre analyse de régression non linéaire :

Erreur standard estimée par le bootstrap. Méthode d'estimation de l'erreur standard d'une statistique par échantillonnage répété de l'ensemble de données d'origine. Pour cela, un échantillonnage (avec remise) est réalisé afin d'obtenir de nombreux échantillons de la même

taille que l'ensemble de données d'origine. Une estimation de l'équation non linéaire est réalisée pour chacun de ces échantillons. L'erreur standard de chaque estimation de paramètres est alors calculée comme l'écart-type estimé par le bootstrap. Les valeurs des paramètres des données d'origine servent de valeurs initiales à chaque échantillon du bootstrap. Cela nécessite un algorithme de programmation quadratique séquentielle.

Méthode d'estimation. Permet de sélectionner la méthode d'estimation, si c'est possible. (Certain choix dans cette boîte de dialogue comme dans d'autres impliquent l'utilisation d'un algorithme de programmation quadratique séquentielle). Les alternatives disponibles sont la programmation quadratique séquentielle et l'algorithme de Levenberg-Marquardt.

- **Programmation quadratique séquentielle..** Cette méthode est utilisable pour des modèles avec ou sans contrainte. La programmation quadratique séquentielle est utilisée automatiquement si vous spécifiez un modèle avec contraintes, une fonction de perte définie par l'utilisateur ou une amorce. Vous pouvez saisir de nouvelles valeurs pour le Maximum d'itérations et la stabilité des coefficients. Vous pouvez également modifier la sélection dans les listes déroulantes de précision à l'optimum, de Précision de la fonction et de critère de convergence.
- **Levenberg-Marquardt.** Algorithme par défaut des modèles non contraints. La méthode Levenberg-Marquardt n'est pas utilisable si vous sélectionnez un modèle avec contraintes, une fonction de perte définie par l'utilisateur ou une amorce. Vous pouvez saisir de nouvelles valeurs pour le Maximum des itérations et vous pouvez également modifier la sélection dans les listes Convergence de la somme des carrés et Convergence des paramètres.

Interpréter les résultats de la régression non linéaire

Les problèmes de régression non linéaire présentent souvent des difficultés de calcul :

- Le choix des valeurs initiales pour les paramètres influence la convergence. Essayez de choisir des valeurs raisonnables et, si possible, proches de la solution finale escomptée.
- Certains algorithmes se révèlent parfois meilleurs que d'autres pour résoudre un problème particulier. Dans la boîte de dialogue Options, sélectionnez l'autre algorithme, le cas échéant. (Si vous indiquez une fonction de perte ou certains types de contrainte, vous ne pouvez pas employer l'algorithme de Levenberg-Marquardt.)
- Lorsque l'itération ne s'interrompt que lorsque le nombre maximal d'itérations est atteint, le modèle final n'est probablement pas une solution satisfaisante. Sélectionnez Utiliser l'analyse précédente pour spécifier les valeurs initiales dans la boîte de dialogue pour poursuivre l'itération ou, encore mieux, choisissez des valeurs initiales différentes.
- Les modèles qui requièrent une mise en exposant de ou par des valeurs importantes peuvent engendrer des dépassements positifs ou négatifs (nombres trop grands ou trop petits pour être représentés sur l'ordinateur). En général, pour éviter cela, vous devez fixer des valeurs initiales appropriées ou fixer des contraintes sur les paramètres.

Fonctions supplémentaires de la commande NLR

Le langage de syntaxe de commande vous permet aussi de :

- Nommer un fichier à partir duquel les valeurs initiales pour les estimations sont lues.

- Spécifier plusieurs déclarations de modèle et fonctions de perte. Cela facilite la spécification d'un modèle segmenté.
- Employer vos propres dérivées plutôt que celles calculées par le programme.
- Spécifier le nombre d'échantillons de départ à générer.
- Indiquez les critères d'itération supplémentaires, notamment la définition d'une valeur critique pour le contrôle de la dérivée et la définition d'un critère de convergence pour la corrélation entre les résidus et les dérivées.

Les critères supplémentaires de la commande `CNLR` (régression non linéaire restreinte) vous permettent de :

- Indiquer le nombre maximal d'itérations mineures permises dans une itération majeure.
- Fixer une valeur critique pour le contrôle de dérivée (calculée).
- Fixer la stabilité des coefficients.
- Indiquer une tolérance pour établir si les valeurs initiales se situent dans les limites déterminées.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
6

Pondération estimée

Les modèles de régression linéaire standard partent du principe que la variance est constante au sein de la population étudiée. En cas contraire (par exemple, lorsque les observations élevées sur un certain attribut montrent plus de variabilité que les observations faibles sur cet attribut), la régression linéaire par la méthode des moindres carrés ordinaires ne fournit plus des estimations optimales. Si les différences de variabilité peuvent être prévues à partir d'une autre variable, la procédure de pondération estimée peut calculer les coefficients d'un modèle de régression linéaire par la méthode des moindres carrés pondérés, de sorte que les observations les plus précises (c'est-à-dire celles offrant le moins de variabilité) ont plus de poids dans la détermination des coefficients de régression. La procédure de pondération estimée teste une fourchette de transformations de la pondération et indique celle qui correspond le mieux aux données.

Exemple : Quels sont les effets de l'inflation et du chômage sur les fluctuations des cours de la bourse ? Les actions à forte valeur montrant plus de variabilité que celles de faible valeur, les moindres carrés ordinaires ne fournissent pas d'estimations optimales. La pondération estimée vous permet de prendre en compte les effets du prix de l'action sur la variabilité des fluctuations de prix dans le calcul du modèle linéaire.

Statistiques. Valeurs de vraisemblance logarithmique de la variable source pondérée testée, *R* multiple, *R* carré, *R* carré ajusté, tableau d'ANOVA pour le modèle WLS, estimations standardisées et non standardisées et vraisemblance logarithmique pour le modèle WLS.

Données : Les variables dépendantes et indépendantes doivent être quantitatives. Les variables qualitatives, comme la religion, la qualification, la zone de résidence, doivent être enregistrées sous forme de variables binaires (factices) ou sous de tout autre type de variables de contraste. La variable de pondération doit être quantitative et doit être associée à la variabilité de la variable dépendante.

Hypothèses : Pour chaque valeur de la variable indépendante, la distribution de la variable dépendante doit être normale. La relation entre la variable dépendante et chaque variable indépendante doit être linéaire et toutes les observations doivent être indépendantes. La variance de la variable dépendante peut varier selon les niveaux de la ou des variables indépendantes mais les différences doivent être prévisibles en fonction de la variable de pondération.

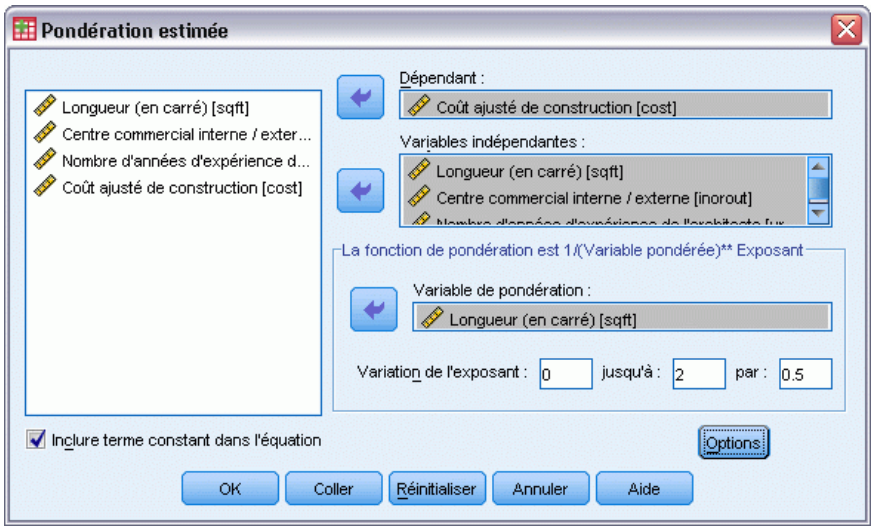
Procédures apparentées : La procédure d'exploration peut être utilisée pour analyser vos données. L'exploration vous propose des tests de normalité et d'homogénéité de la variance, ainsi que des illustrations graphiques. Si votre variable dépendante semble avoir la même variance sur tous les niveaux des variables indépendantes, utilisez la procédure de régression linéaire. Si vos données apparaissent ne pas satisfaire une hypothèse (telle que la normalité), essayez de les modifier. Si vos données ne sont pas liées linéairement et qu'une modification ne change rien, utilisez un autre modèle dans la procédure d'ajustement des fonctions. Si votre variable dépendante est dichotomique, telle que Utilisable ou Défectueux, utilisez la procédure de régression logistique.

Si votre variable dépendante est censurée (par exemple, la durée de survie après opération), utilisez les procédures Durée de vie, Kaplan-Meier ou Régression de Cox, disponibles dans l'option Statistiques avancées. Si vos données ne sont pas indépendantes (par exemple, si vous observez le même individu sous différentes conditions), utilisez la procédure de mesures répétées, dans l'option Statistiques avancées.

Obtenir une analyse de pondération estimée

▶ A partir des menus, sélectionnez :
Analyse > Régression > Pondération estimée...

Figure 6-1
Boîte de dialogue Pondération estimée



▶ Sélectionnez une variable dépendante.

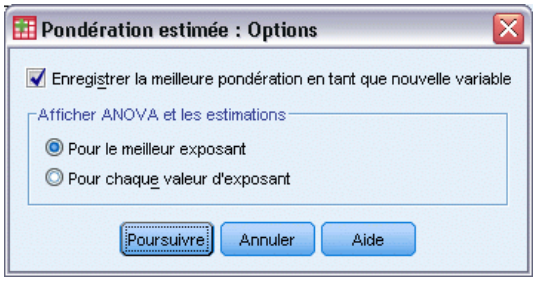
▶ Sélectionnez une ou plusieurs variables indépendantes.

▶ Sélectionnez la variable qui est la source de l'hétéroscédasticité comme variable de pondération.

- **Variable de pondération.** Les données sont pondérées par la réciproque de cette variable élevée à une puissance. L'équation de régression est calculée pour chacune des valeurs d'un intervalle spécifié d'exposants et indique l'exposant qui maximise la fonction de log-vraisemblance.
- **Variation de l'exposant.** S'utilise de pair avec la variable de pondération pour calculer les pondérations. Plusieurs équations de régression seront acceptables, une par valeur de l'intervalle d'exposants. Les valeurs indiquées dans la zone de test de variation d'exposant et dans la zone de texte doivent être comprises entre 6,5 et 7,5 (limites incluses). Les valeurs d'exposant varient de la plus faible à la plus élevée, l'incrément étant déterminé par la valeur spécifiée. Le nombre total de valeurs dans la variation de l'exposant est limité à 150.

Options de la pondération estimée

Figure 6-2
Boîte de dialogue Pondération estimée : Options



Vous pouvez sélectionner les options de votre analyse de pondération estimée :

Enregistrer meilleure pondération dans nouvelle variable. Ajoute la variable de pondération au fichier actif. Cette variable s'appelle *WGT_n*, *n* étant le nombre attribué à la variable pour qu'elle ait un nom univoque.

Afficher ANOVA et les estimations. Vous permet de contrôler le mode d'affichage des statistiques dans le résultat. Vous pouvez choisir entre Pour le meilleur exposant et Pour chaque valeur de l'exposant.

Fonctions supplémentaires de la commande WLS

Le langage de syntaxe de commande vous permet aussi de :

- Fournir une valeur unique à l'exposant.
- Spécifier la liste des valeurs d'exposant ou combiner un intervalle de valeurs avec une liste de valeurs pour cet exposant.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
7

Régression par les doubles moindres carrés

Les modèles de régression linéaire standard partent du principe que les erreurs au niveau de la variable dépendante ne sont pas corrélées à la ou les variables indépendantes. En cas contraire (par exemple, lorsque les relations entre les variables sont bidirectionnelles), la régression linéaire par la méthode des moindres carrés ne constitue plus un modèle de prévision optimal. La régression par les doubles moindres carrés emploie des variables instrumentales non corrélées aux termes d'erreurs pour calculer les valeurs prévisionnelles du prédicteur problématique (première étape) puis utilise ces valeurs calculées pour évaluer le modèle de régression linéaire de la variable dépendante (seconde étape). Les valeurs calculées étant fondées sur des variables non corrélées aux erreurs, les résultats du modèle double sont optimaux.

Exemple : La demande pour un article est-elle liée au prix et au revenu du consommateur ? La difficulté ici réside dans le fait que le prix et la demande agissent mutuellement l'un sur l'autre. En effet, le prix influence la demande mais la demande influence également le prix. Un modèle de régression par les doubles moindres carrés peut utiliser le revenu du consommateur comme un représentant du prix qui n'est pas corrélé avec les erreurs de mesure de la demande. Ce représentant joue le rôle du prix lui-même dans le modèle originalement spécifié, celui-ci est alors évalué.

Statistiques : Pour chaque modèle : coefficients de régression standardisés et non standardisés, *R* multiple, R^2, R^2 ajusté, erreur standard de la prévision, tableau d'analyse de la variance, prévisions, résidus et intervalles de prévision. Egalement, intervalles de confiance de 95 % pour chaque coefficient de régression, matrices de corrélation et de covariance des estimations des paramètres.

Données : Les variables dépendantes et indépendantes doivent être quantitatives. Les variables qualitatives, comme la religion, la qualification, la zone de résidence, doivent être enregistrées sous forme de variables binaires (factices) ou sous de tout autre type de variables de contraste. Les variables explicatives **endogènes** doivent également être quantitatives (pas catégorielles).

Hypothèses : Pour chaque valeur de la variable indépendante, la distribution de la variable dépendante doit être normale. La variance de la distribution de la variable dépendante doit être constante pour toutes les valeurs de la variable indépendante. La relation entre la variable dépendante et chaque variable indépendante doit être linéaire.

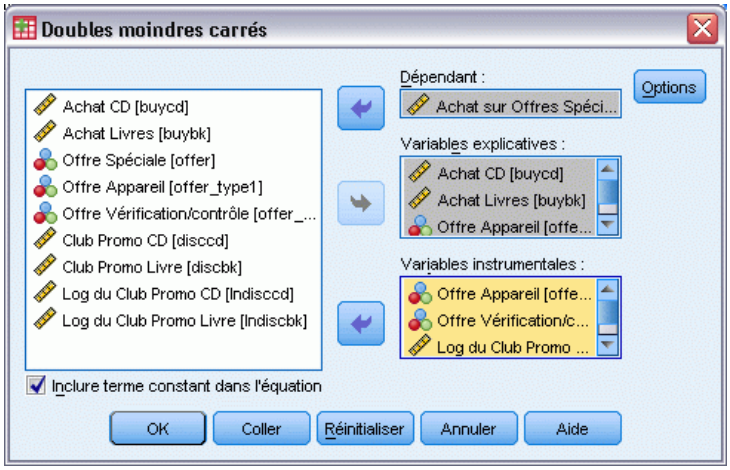
Procédures apparentées : Si vous estimez qu'aucune de vos variables explicatives n'est corrélée avec les erreurs de la variable dépendante, vous pouvez employer une procédure de régression linéaire. Si vos données ne semblent pas répondre aux hypothèses formulées (telles que la normalité et la constance de la variance), essayez de les modifier. Si vos données ne sont pas liées linéairement et qu'une modification ne change rien, utilisez un autre modèle dans la procédure d'ajustement des fonctions. Si votre variable dépendante est dichotomique, telle que

Vendue ou Non vendue, utilisez la procédure de régression logistique. Si vos données ne sont pas indépendantes (par exemple, si vous observez le même individu sous différentes conditions), utilisez la procédure de mesures répétées, dans l'option Statistiques avancées.

Obtenir une analyse de la régression par les doubles moindres carrés

▶ A partir des menus, sélectionnez :
Analyse > Régression > Doubles moindres carrés...

Figure 7-1
Boîte de dialogue Doubles moindres carrés



▶ Sélectionnez une variable dépendante.

▶ Sélectionnez une ou plusieurs Variables explicatives.

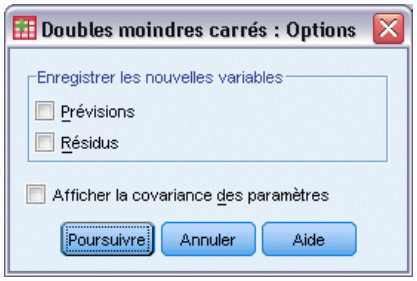
▶ Sélectionnez une ou plusieurs Variables instrumentales.

- **Variables instrumentales.** Ce sont les variables utilisées pour calculer les prévisions des variables endogènes dans la première phase de l'analyse double moindres carrés. Les mêmes variables peuvent apparaître à la fois dans les zones de liste Variables explicatives et Variables instrumentales. Le nombre de variables instrumentales doit être au moins aussi élevé que celui des variables explicatives. Si toutes les variables explicatives et instrumentales répertoriées sont identiques, les résultats sont les mêmes que ceux obtenus par la procédure de régression linéaire.

Les variables explicatives non spécifiées comme instrumentales sont considérées comme étant endogènes. En principe, toutes les variables exogènes de la liste Explicatif sont également spécifiées en tant que variables instrumentales.

Options de régression par les doubles moindres carrés

Figure 7-2
Boîte de dialogue Doubles moindres carrés : Options



Vous pouvez sélectionner les options suivantes pour votre analyse :

Enregistrer les nouvelles variables. Permet d'ajouter de nouvelles variables au fichier actif. Les options disponibles sont Prévisions et Résidus.

Afficher la covariance des paramètres. Permet d'imprimer la matrice de covariance des estimations des paramètres.

Fonctions supplémentaires de la commande 2SLS

Le langage de syntaxe de commande vous permet également d'estimer plusieurs équations en même temps. Pour obtenir des informations complètes sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Annexe

A

Méthodes de codification des variables qualitatives

Dans de nombreuses procédures, vous pouvez demander le remplacement automatique d'une variable indépendante qualitative par un ensemble de variables de contraste, qui seront ensuite introduites dans une équation, ou en seront supprimées, en tant que bloc. Vous pouvez indiquer comment le groupe de variables de contraste doit être codé, généralement à l'aide de la sous-commande CONTRAST. Cette annexe explique et illustre le fonctionnement des différents types de contraste que vous pouvez appeler via la sous-commande CONTRAST.

Écart type

Ecart par rapport à la moyenne générale. Dans les matrices, ces contrastes ont la forme suivante :

moyenne	(1/*k*	1/*k*	···	1/*k*	1/*k*)
ddl(1)	(1–1/*k*	–1/*k*	···	–1/*k*	–1/*k*)
ddl(2)	(–1/*k*	1–1/*k*	···	–1/*k*	–1/*k*)
.		.			
.		.			
df(k–1)	(–1/*k*	–1/*k*	···	1–1/*k*	–1/*k*)

où *k* est le nombre de modalités de la variable indépendante, la dernière modalité étant omise par défaut. Par exemple, les contrastes d'écart d'une variable indépendante comportant trois modalités sont les suivants :

(1/3	1/3	1/3)
(2/3	–1/3	–1/3)
(–1/3	2/3	–1/3)

Pour omettre une modalité autre que la dernière, indiquez son numéro entre parenthèses après le mot-clé DEVIATION. Par exemple, la sous-commande suivante permet d'obtenir les écarts de la première et de la troisième modalité, et d'omettre la deuxième :

```
/CONTRAST(FACTOR)=DEVIATION(2)
```

Supposons que le facteur (*FACTOR*) comporte trois modalités. La matrice de contraste obtenue est la suivante :

(1/3	1/3	1/3)
(2/3	–1/3	–1/3)
(–1/3	–1/3	2/3)

Simple

Contrastes simples. Compare chaque niveau d'un facteur au dernier. La forme de la matrice générale est la suivante :

moyenne	(1/*k*	1/*k*	...	1/*k*	1/*k*)
ddl(1)	(1	0	...	0	–1)
ddl(2)	(0	1	...	0	–1)
.		.			
.		.			
df(k–1)	(0	0	...	1	–1)

où *k* est le nombre de modalités de la variable indépendante. Par exemple, les contrastes simples d'une variable indépendante comportant quatre modalités sont les suivants :

(1/4	1/4	1/4	1/4)
(1	0	0	–1)
(0	1	0	–1)
(0	0	1	–1)

Pour utiliser comme modalité de référence une autre modalité que la dernière, indiquez entre parenthèses, après le mot-clé `SIMPLE`, le numéro de séquence de la modalité de référence ; il ne s'agit pas nécessairement de la valeur associée à la modalité. Par exemple, la sous-commande `CONTRAST` suivante permet d'obtenir une matrice de contraste qui omet la deuxième modalité :

```
/CONTRAST(FACTOR) = SIMPLE(2)
```

Supposons que le facteur (*FACTOR*) comporte quatre modalités. La matrice de contraste obtenue est la suivante :

(1/4	1/4	1/4	1/4)
(1	–1	0	0)
(0	–1	1	0)
(0	–1	0	1)

Helmert

Contrastes de Helmert. Compare les modalités d'une variable indépendante avec la moyenne des modalités suivantes. La forme de la matrice générale est la suivante :

moyenne	(1/*k*	1/*k*	···	1/*k*	1/*k*)
ddl(1)	(1	–1/(*k*–1)	···	–1/(*k*–1)	–1/(*k*–1))
ddl(2)	(0	1	···	–1/(*k*–2)	–1/(*k*–2))
.		.			
.		.			
df(k–2)	(0	0	1	–1/2	–1/2
df(k–1)	(0	0	···	1	–1)

où *k* est le nombre de modalités de la variable indépendante. Par exemple, une variable indépendante comportant quatre modalités présente une matrice de contraste de Helmert ayant la forme suivante :

(1/4	1/4	1/4	1/4)
(1	–1/3	–1/3	–1/3)
(0	1	–1/2	–1/2)
(0	0	1	–1)

Différencié d'ordre

Contrastes de différence ou contrastes inversés de Helmert. Compare les modalités d'une variable indépendante avec la moyenne des modalités précédentes de la variable. La forme de la matrice générale est la suivante :

moyenne	(1/*k*	1/*k*	1/*k*	···	1/*k*)
ddl(1)	(–1	1	0	···	0)
ddl(2)	(–1/2	–1/2	1	···	0)
.		.			
.		.			
df(k–1)	(–1/(*k*–1)	–1/(*k*–1)	–1/(*k*–1)	···	1)

où *k* est le nombre de modalités de la variable indépendante. Par exemple, les contrastes de différence d'une variable indépendante comportant quatre modalités sont les suivants :

(1/4	1/4	1/4	1/4)
(–1	1	0	0)
(–1/2	–1/2	1	0)
(–1/3	–1/3	–1/3	1)

Polynomial

Contraste polynomial orthogonal. Le premier degré de liberté contient l'effet linéaire sur toutes les modalités, le second degré l'effet quadratique, le troisième degré l'effet cubique, et ainsi de suite pour les effets d'ordre supérieur.

Vous pouvez définir l'espacement entre les seuils du traitement mesuré par la variable qualitative donnée. Vous pouvez indiquer l'espacement égal (espacement par défaut en cas d'omission de la mesure), sous la forme d'une suite d'entiers allant de 1 à *k*, où *k* est le nombre de modalités. Si la variable *médicament* comporte trois modalités, la sous-commande

```
/CONTRAST(DRUG)=POLYNOMIAL
```

est identique à

```
/CONTRAST(DRUG)=POLYNOMIAL(1,20,3)
```

Toutefois, l'espacement égal n'est pas systématiquement nécessaire. Par exemple, supposons que la variable *médicament* représente différents dosages d'un médicament administré à trois groupes. Si le dosage administré au deuxième groupe est le double de celui administré au premier groupe, et que celui administré au troisième groupe est le triple de celui administré au premier groupe, les modalités de traitement sont espacées de manière égale et, dans cette situation, une mesure appropriée se compose d'une suite d'entiers :

```
/CONTRAST(DRUG)=POLYNOMIAL(1,20,3)
```

Toutefois, si le dosage administré au deuxième groupe est le quadruple de celui administré au premier groupe, et que celui administré au troisième groupe est le septuple de celui administré au premier groupe, une mesure appropriée se présente sous la forme suivante :

```
/CONTRAST(DRUG)=POLYNOMIAL(1,4,7)
```

Dans les deux cas, une fois le contraste défini, le premier degré de liberté de la variable *médicament* contient l'effet linéaire des niveaux de dosage, tandis que le deuxième degré contient l'effet quadratique.

Les contrastes polynomiaux sont particulièrement utiles pour réaliser des tests de tendances et analyser la nature des surfaces de réponses. Vous pouvez également utiliser les contrastes polynomiaux pour effectuer un ajustement de courbe non linéaire, comme une régression curviligne.

Répété

Compare les seuils adjacents d'une variable indépendante. La forme de la matrice générale est la suivante :

moyenne	(1/*k*	1/*k*	1/*k*	···	1/*k*	1/*k*)
ddl(1)	(1	–1	0	···	0	0)
ddl(2)	(0	1	–1	···	0	0)
.		.				

	.		.			
df(k–1)	(0	0	0	···	1	–1)

où *k* est le nombre de modalités de la variable indépendante. Par exemple, les contrastes répétés d'une variable indépendante comportant quatre modalités sont les suivants :

(1/4	1/4	1/4	1/4)
(1	–1	0	0)
(0	1	–1	0)
(0	0	1	–1)

Ces contrastes sont utiles dans l'analyse des profils et lorsque des statistiques de différence sont nécessaires.

Spécial

Contraste défini par l'utilisateur. Permet la saisie de contrastes spéciaux sous la forme de matrices carrées comportant autant de lignes et de colonnes que le nombre de modalités de la variable indépendante spécifiée. Pour MANOVA et LOGLINEAR, la première ligne saisie est toujours l'effet de moyenne ou de constante, et représente le groupe de pondérations indiquant comment déterminer, par rapport à la variable spécifiée, la moyenne des autres variables indépendantes (le cas échéant). Généralement, ce contraste est un vecteur.

Les autres lignes de la matrice contiennent les contrastes spéciaux indiquant les comparaisons souhaitées entre les modalités de la variable. Généralement, les contrastes orthogonaux sont les plus utiles. Ils ne sont pas redondants et sont statistiquement indépendants. Les contrastes sont orthogonaux si :

- Pour chaque ligne, la somme des coefficients de contraste est égale à 0.
- La somme des produits des coefficients correspondant à toutes les paires de lignes disjointes est aussi égale à 0.

Par exemple, supposons que la variable traitement comporte quatre niveaux et que vous souhaitez comparer les différents seuils de traitement. Un contraste spécial approprié peut avoir la forme suivante :

(1	1	1	1)	pondérations pour le calcul de la moyenne
(3	–1	–1	–1)	comparaison du premier seuil avec les deuxième, troisième et quatrième
(0	2	–1	–1)	comparaison du deuxième seuil avec les troisième et quatrième
(0	0	1	–1)	comparaison des troisième et quatrième seuils

que vous définissez à l'aide de la sous-commande CONTRAST suivante pour MANOVA, LOGISTIC REGRESSION et COXREG :

```
/CONTRAST(TREATMNT)=SPECIAL( 1  1  1  1
                             3 -1 -1 -1
                             0  2 -1 -1
```

```
                                    0   0   1 -1 )
```

Pour LOGLINEAR, vous devez indiquer :

```
/CONTRAST(TREATMNT)=BASIS SPECIAL( 1   1   1   1
                                   3  -1  -1  -1
                                   0   2  -1  -1
                                   0   0   1  -1 )
```

La somme de chaque ligne, à l'exception de la ligne des moyennes, est égale à 0, de même que celle des produits de chaque paire de lignes disjointes :

Lignes 2 et 3 :	(3)(0) + (–1)(2) + (–1)(–1) + (–1)(–1) = 0
Lignes 2 et 4 :	(3)(0) + (–1)(0) + (–1)(1) + (–1)(–1) = 0
Lignes 3 et 4 :	(0)(0) + (2)(0) + (–1)(1) + (–1)(–1) = 0

Il n'est pas nécessaire que les contrastes spéciaux soient orthogonaux. Toutefois, ils ne doivent pas constituer des combinaisons linéaires les uns avec les autres. Si tel est le cas, la procédure signale la dépendance linéaire et interrompt le traitement. Les contrastes polynomiaux, de différence et de Helmert sont tous des contrastes orthogonaux.

Indicateur

Codage des variables indicatrices. Egalement appelé codage de façon fictive, ce type de codage n'est pas disponible dans LOGLINEAR et MANOVA. Le numéro des nouvelles variables codées est *k*–1. Les observations de la modalité de référence sont codées 0 pour toutes les variables *k*–1. Une observation dans la $n^{\text{ième}}$ modalité est codée 0 pour toutes les variables indicatrices, sauf la $n^{\text{ième}}$, codée 1.

Annexe
B

Notices

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