

# IBM SPSS Missing Values 21



注：この情報とサポートされている製品をご使用になる前に、「注意事項」（p.103）の一般情報をお読みください。

本版は IBM® SPSS® Statistics 21 ,および新版で指示されるまで後続するすべてのリリースおよび変更に対して適用されます。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

Licensed Materials - Property of IBM

© Copyright IBM Corporation 1989, 2012.

U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

---

# はじめに

IBM® SPSS® Statistics は、データ分析の包括的システムです。欠損値は、このマニュアルで説明されている追加の分析手法を提供するオプションのアドオン モジュールです。欠損値 アドオン モジュールは SPSS Statistics Core システムと組み合わせて使用し、Core システムに完全に統合されます。

## IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス パフォーマンスを向上させるために信頼する完全で、一貫した正確な情報を提供します。ビジネス インテリジェンス、予測分析、財務実績および戦略管理、および 分析アプリケーションの包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な業界のソリューション、実績ある実例、専門サービスと組み合わせ、さまざまな規模の組織が、高い生産性を実現、意思決定を自信を持って自動化し、より良い決定をもたらします。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。IBM SPSS ソフトウェアを日々の業務に取り入れることによって、組織は業務目標を達成し、大きな競争的優位を獲得することができるよう、意思決定を方向付け、自動化することができるようになります。お問い合わせは、<http://www.ibm.com/spss> を参照してください。

## テクニカル サポート

テクニカル サポートのサービスをご利用いただけます。IBM Corp. 製品の使用方法や、対応しているハードウェア環境へのインストールに関して問い合わせることもできます。テクニカル サポートの詳細については、IBM Corp. Web サイト (<http://www.ibm.com/support>) を参照してください。連絡の際は、所属団体名、サポート契約などを確認できるよう、あらかじめ手元にご用意ください。

## 学生向けテクニカル サポート

IBM SPSS ソフトウェア製品の Student 版、アカデミック版、Grad パック版を使用している学生の場合、学生用の特別オンライン ページ、[Solutions for Education \(http://www.ibm.com/spss/rd/students/\)](http://www.ibm.com/spss/rd/students/) ページを参照してください。大学提供の IBM SPSS ソフトウェアのコピーを使用している場合、大学の IBM SPSS 製品コーディネータにお問い合わせください。

## カスタマ サービス

配送やアカウントに関するご質問は、お近くの営業所にお問い合わせください。お問い合わせの際には、シリアル番号をご用意ください。

## トレーニング セミナー

IBM Corp. では一般公開およびオンサイトで トレーニング セミナーを実施しています。セミナーでは実践的な講習を行います。セミナーは主要都市で定期的開催されます。セミナーに関する詳細については、<http://www.ibm.com/software/analytics/spss/training> を参照してください。

---

# 内容

## パート I: ユーザー ガイド

<b>1</b>	<b>欠損値の概要</b>	<b>1</b>
<b>2</b>	<b>欠損値分析</b>	<b>2</b>
	欠損値のパターンの表示 . . . . .	5
	欠損値の記述統計の表示 . . . . .	7
	統計量の推定と欠損値の代入 . . . . .	8
	EM 推定のオプション . . . . .	10
	回帰推定のオプション . . . . .	11
	予測される変数と予測する変数 . . . . .	12
	MVA コマンドの追加機能 . . . . .	13
<b>3</b>	<b>多重代入</b>	<b>15</b>
	パターン分析 . . . . .	17
	欠損データ値の代入 . . . . .	18
	方法 . . . . .	21
	制約 . . . . .	23
	出力 . . . . .	26
	MULTIPLE IMPUTATION コマンドの追加機能 . . . . .	27
	多重代入データを使用した作業 . . . . .	27
	多重代入データの分析 . . . . .	31
	多重代入オプション . . . . .	38

## パート II: 例

<b>4</b>	<b>欠損値分析</b>	<b>41</b>
	欠損データのパターンの記述 . . . . .	41

分析を実行して記述統計量を表示 . . . . .	41
記述統計の評価 . . . . .	43
パターンを表示するために分析を再実行 . . . . .	50
パターン テーブルの評価 . . . . .	52
Little の MCAR 検定のために分析を再実行 . . . . .	53

## 5 [多重代入] 55

多重代入を使用したデータセットの比較と分析 . . . . .	55
欠損値のパターンの分析 . . . . .	55
欠損値の自動代入 . . . . .	60
ユーザー指定による代入モデル . . . . .	67
FCS 収束の確認 . . . . .	76
完成したデータの分析 . . . . .	80
[要約] . . . . .	91

## 付録

### A サンプル ファイル 92

### B 注意事項 103

## 索引 106

# パート I: ユーザー ガイド





# 欠損値の概要

典型的なモデリング手順は単に欠損値を伴うケースを分析から破棄するため、これらのケースにより難題がもたらされます。いくつかの欠損値（すべてのケースのうちのおよそ 5% 未満）があり、それらの値がランダムに欠損している（つまり値が欠損しているかどうかは他の値に依存しない）場合、典型的なリストごとの削除の方法は比較的“安全”です。欠損値オプションは、リストごとの削除が十分かどうか判断する際に役立ちます。また十分でない場合は、このオプションにより欠損値を取り扱う方法が提供されます。

## 欠損値分析対多重代入手順

欠損値オプションは、欠損値を取り扱う手順のセットを 2 つ提供します。

- **多重代入**手順は、最終的な欠損値の多重代入に適合された、欠損データのパターン分析を提供します。つまり、データセットの多重バージョンが作成され、それぞれのバージョンには独自の代入値のセットが含まれます。統計分析が実行される際、すべての代入されたデータセットに対するパラメータ推定値はプールされます。この手順では、一般的に単独の代入より正確な推定値が提供されます。
- **欠損値分析**は、欠損データ分析用の記述ツール（特に Little の MCAR 検定）の微妙に異なるセットを提供します。この分析にはさまざまな単一代入方法が含まれます。通常、多重代入は単一代入よりも優れていると考えられている点に注意してください。

## 欠損値タスク

以下の基本的な手順に従って、欠損値の分析を行えます。

- ▶ **欠損を調べる。** 欠損値分析とパターン分析を使用してデータ中の欠損値のパターンを調査して、多重代入が必要かどうか判断します。
- ▶ **欠損値を代入(I)** 欠損データ値の代入を使用して、欠損値を多重代入します。
- ▶ **“完成した”データを分析する。** 多重代入データをサポートする任意の手順を使用する。多重代入データセットの分析およびこれらのデータをサポートする手順のリストの詳細については、[多重代入データの分析](#) p. 31 を参照してください。

# 欠損値分析

Missing Value Analysis 手続きでは、次の 3つの主要な機能を実行します。

- 欠損データのパターンを示す。欠損値のある場所。欠損値の程度。変数ペアには、複数のケースにおいて値が欠損する傾向があるかどうか。データ値は極値であるかどうか。値はランダムに欠損しているか。
- リストごと、ペアごと、回帰法、または EM（期待値の最大化）法といたさまざまな欠損値手法で平均値、標準偏差、共分散、および相関を推定する。ペアごとの手法では、ペアごとの完全なケースの度数も表示されます。
- 回帰法または EM 法を使用して欠損値に推定値を代入します。しかし、多重代入のほうが、より正確な結果を通常生み出すと考えられています。

欠損値分析では、不完全データが原因であるいくつかの懸念に対応できません。欠損値のあるケースが欠損値のないケースと系統的に異なる場合、その結果が誤解を招くものとなることがあります。また、情報が最初の計画より少ないため、欠損データによって、計算された統計量の精度が低下する場合があります。もう 1 つの懸念は、多くの統計手続きにおける仮定が完全なケースに基づいており、欠損値があると必要な理論が複雑になる場合があるということです。

**例:** 白血病の治療法を評価する際に、数個の変数を測定するとします。ただし、全患者の測定値がすべて揃っているわけではありません。欠損データのパターンを表示、集計した結果、ランダムであることがわかりました。EM 分析を使用して、平均値、相関、および共分散を推定します。これは、データが完全にランダムに欠損しているかどうかを判断するためにも使われます。次に、欠損値は代入値に置き換えられ、さらに分析するために新しいデータファイルに保存されます。

**統計量。** 非欠損値数、平均値、標準偏差、欠損値数、極値数などの 1 変数統計量。リストごと、ペアごと、EM 法、または回帰法を使用した推定平均値、分散共分散行列、相関行列。EM 結果による Little の MCAR 検定。手法別の平均集計。欠損値と非欠損値で定義されるグループの場合は t 検定。すべての変数： ケース対変数に表示される欠損値パターン。

## データの考慮事項

**データ。** データは、カテゴリ変数または量的変数（スケール変数または連続変数）のどちらでもかまいません。ただし、統計量を推定して、量的変数の場合にのみ欠損値を代入することができます。各変数では、システム欠

損値としてコード化されていない欠損値は、ユーザー欠損値として定義する必要があります。たとえば、あるアンケート項目に対し、「分からない」という回答が 5 でコード化されており、これを欠損値として扱いたい場合は、5 をユーザー欠損値としてコード化する必要があります。

**度数による重み付け。**度数（複製）による重み付けは、この手続きによって処理されます。負の値または 0 の複製による重み付けを持つケースは無視されます。非整数値の重み付けは切り捨てられます。

**仮定。**リストごと、ペアごと、および回帰の推定では、欠損値パターンがデータ値に依存しないと仮定します（この条件を**完全にランダムな欠損**、または MCAR と呼びます）。したがって、データが MCAR であるとき、すべての推定手法（EM 法も含む）では相関および共分散の一貫した不偏推定値が得られます。MCAR 仮定に反すると、リストごと、ペアごと、および回帰法によって得られる推定値が偏ることがあります。データが MCAR ではない場合は、EM 推定を使用する必要があります。

EM 推定は、欠損値のパターンは観測データのみに関連するという仮定に依存します（この条件を**ランダムな欠損**、または MAR と呼びます）。この仮定では、利用可能な情報を使用して推定値を調整できます。たとえば、教育と収入の調査において、教育レベルの低い被験者のほうが収入の欠損値が多くなる可能性があります。このケースでは、データは MCAR ではなく MAR です。つまり、MAR の場合、収入が記録される確率は被験者の教育レベルに依存します。この確率は、その教育レベル内の収入ではなく、教育によって異なります。また収入が記録される確率が各教育レベル内の収入の値によっても異なる場合（たとえば、高収入の人々は収入を報告しないなど）、データは MCAR でも MAR でもありません。この状況は一般的ではありませんが、該当する場合、どちらの手法も適切ではありません。

**関連手続き。**多くの手続きでは、リストごとまたはペアごとの推定を使用することができます。線型回帰および因子分析では、平均値で欠損値を置き換えることができます。予測アドオン モジュールには、時系列の欠損値を置き換える複数の手法が用意されています。

### 欠損値分析を行うには

- ▶ メニューから次の項目を選択します。  
分析(A) > 欠損値分析...

図 2-1  
[欠損値分析] ダイアログ ボックス



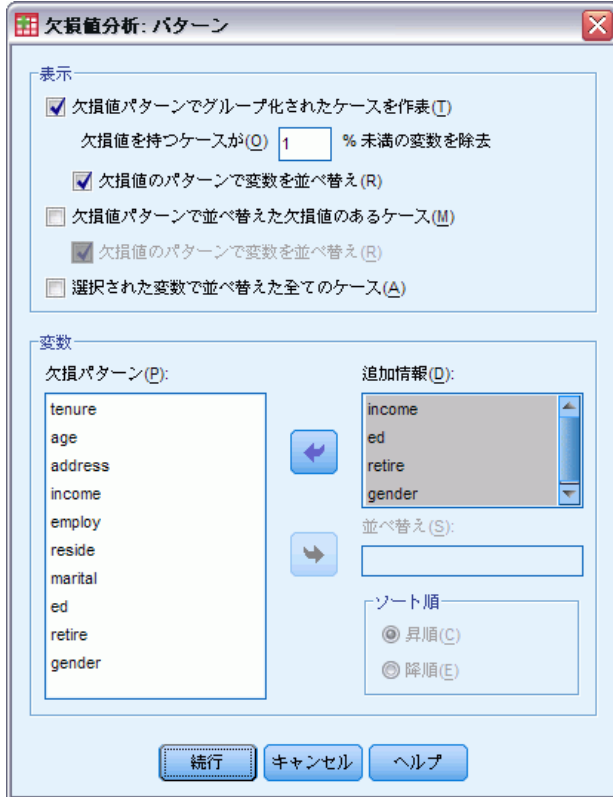
- ▶ 統計量を推定し、オプションで欠損値を代入するために、量的（スケール）変数を選択します。

オプションとして、次の選択が可能です。

- カテゴリ変数（数値型または文字型）を選択し、カテゴリ数の上限（[Maximum Categories (最大カテゴリ数)]）を入力します。
- [パターン] をクリックすると、欠損値のパターンが集計されます。詳細は、[p. 5 欠損値のパターンの表示](#) を参照してください。
- [記述統計] をクリックすると、欠損値の記述統計が表示されます。詳細は、[p. 7 欠損値の記述統計の表示](#) を参照してください。
- 統計の推定手法（平均、共分散、相関）と代入する欠損値を選択します。詳細は、[p. 8 統計量の推定と欠損値の代入](#) を参照してください。
- EM または回帰を選択する場合は、[変数] をクリックして推定に使用するサブセットを指定します。詳細は、[p. 12 予測される変数と予測する変数](#) を参照してください。
- ケースのラベル変数を選択します。この変数は、個々のケースを表示するパターン テーブル内のケースにラベルを付けるために使用します。

## 欠損値のパターンの表示

図 2-2  
[Missing Value Analysis Patterns (欠損値分析 - パターン)] ダイアログ ボックス



欠損値のパターンと程度を示すさまざまなテーブルを表示できます。これらのテーブルを使って、次を確認できます。

- 欠損値のある場所
- 変数ペアが個々のケースで値を欠損する傾向
- データ値が極値かどうか

### 表示

欠損データのパターンを表示するために、3種類のテーブルが用意されています。

**ケースの集計。** 分析変数の欠損値パターンが集計され、各パターンの度数分布表が表示されます。パターンの類似性によって度数と変数を並べ替える場合は、[Sort variables by missing value pattern (欠損値パターンごとに変数を並べ替える)]を使用します。発生日数の低いパターンを除外する場合は、[Omit patterns with less than n % of cases (n % 未満のケースのパターンを省略する)]を使用します。

**欠損値のあるケース。** 欠損値または極値のある各ケースが、各分析変数に対して作表されます。パターンの類似性によって度数と変数を並べ替える場合は、[Sort variables by missing value pattern (欠損値パターンごとに変数を並べ替える)] を使用します。

**すべてのケース。** 各ケースが作表され、各変数の欠損値および極値が示されます。[並べ替え] で変数を指定しない限り、ケースはデータファイルでの表示順と同じ順序で一覧表示されます。

個々のケースを表示するテーブルでは、次の記号が使用されます。

+	極大値
-	極小値
S	システム欠損値
A	ユーザー欠損値の第 1 タイプ
B	ユーザー欠損値の第 2 タイプ
C	ユーザー欠損値の第 3 タイプ

## 変数

分析に含まれている変数についての追加情報を表示できます。[Additional Information for (追加情報)] に追加した変数は、欠損パターン テーブルに個々に表示されます。量的 (スケール) 変数の場合は平均が表示され、カテゴリ変数の場合は各カテゴリのパターンがあるケースの数が表示されます。

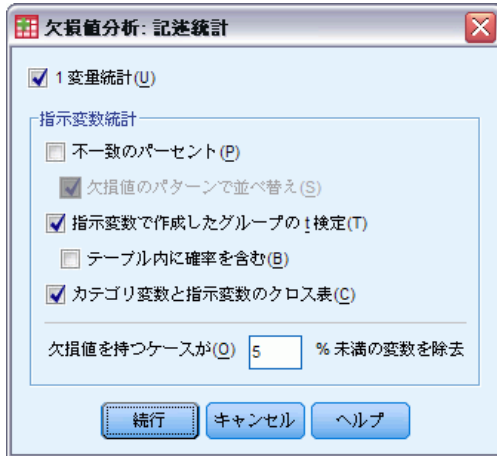
- **並べ替え。** 指定した変数のケースが昇順または降順で並べ替えられます。[全てのケース] でのみ使用できます。

## 欠損値パターンを表示するには

- ▶ [欠損値分析] メイン ダイアログ ボックスで、欠損値パターンを表示する変数を選択します。
- ▶ [パターン] をクリックします。
- ▶ 表示するパターン テーブルを選択します。

## 欠損値の記述統計の表示

図 2-3  
[欠損値分析 - 記述統計] ダイアログ ボックス



### 1 変量統計

1 変量統計は、欠損値の一般的な程度を特定するのに役立ちます。各変数について次が表示されます。

- 非欠損値の数
- 欠損値の数とパーセント

量的（スケール）変数の場合は、次も表示されます。

- 平均値
- 標準偏差(D)
- 極大値と極小値の数

### 指示変数統計

各変数に指示変数が作成されます。このカテゴリ変数は、個々のケースに変数があるかないかを示します。指示変数は、不一致、t 検定、および度数分布表を作成するために使用します。

**不一致のパーセント。** 変数のペアごとに、一方の変数は欠損値、もう一方の変数は非欠損値というケースのパーセントを表示します。表中の各対角要素には、1 つの変数の欠損値のパーセントが含まれます。

**指示変数ごとに形成されたグループによる t 検定。** スチューデントの T 統計量を使って、各量的変数の 2 グループの平均値を比較します。グループによって、変数があるかないかを指定します。2 つのグループの t 統計量、自由度、欠損値および非欠損値の度数、および平均が表示されます。ま

た、t 統計量に関する両側確率も表示できます。分析で複数の検定を行う場合、これらの確率を有意性検定に使用しないでください。これらの確率は、1 つの検定を行う場合にのみ適しています。

**カテゴリおよび指示変数のクロス集計表。** カテゴリ変数ごとの表が表示されます。それぞれのカテゴリについて、他の変数の非欠損値の度数とパーセントがこの表に表示されます。各欠損値タイプのパーセントも表示されます。

**n % 未満のケースの欠損変数を省略する。** 表を小さくするために、少数のケースで計算される統計量を除外することができます。

#### 記述統計を表示するには

- ▶ [欠損値分析] メイン ダイアログ ボックスで、欠損値の記述統計量を表示する変数を選択します。
- ▶ [記述統計] をクリックします。
- ▶ 表示する記述統計を選択します。

## 統計量の推定と欠損値の代入

リストごと（完全なケースのみ）、ペアごと、EM（期待値の最大化）法、または回帰法を使用して、推定平均、標準偏差、共分散、相関を推定できます。また、欠損値を代入することもできます（推定置換値）。欠損値の問題の解決には、通常、**多重代入**が単一代入よりも優れていると考えられている点に注意してください。代入が必要かどうか判断するには、Little の MCAR 検定が便利です。

#### リストごと

この手法では、完全なケースのみを使用します。いずれかの分析変数に欠損値が含まれている場合、そのケースは計算から除外されます。

#### ペアごと

この手法では、ペアの分析変数を調べ、両方の変数に非欠損値が含まれている場合にのみケースを使用します。度数、平均、標準偏差は、ペアごとに個別に計算されます。ケース中の他の欠損値は無視されるため、2 つの変数の相関と共分散はその他の変数の欠損値には依存しません。

#### EM 法

この手法では、一部欠損データの分布を仮定し、この分布の尤度を基に推論を導き出します。各反復は、E ステップと M ステップで構成されています。E ステップでは、パラメータの観測値と現在の推定値を前提に、「欠



損」データの条件付き期待値を求めます。次に、これらの期待値は「欠損」データと置き換えられます。M ステップでは、欠損データが入力されている場合でも、パラメータの最大尤度推定が計算されます。欠損値が直接入力されていないため、「欠損」のようにかっこで囲みます。その代わり、これらの関数は対数尤度で使用されます。

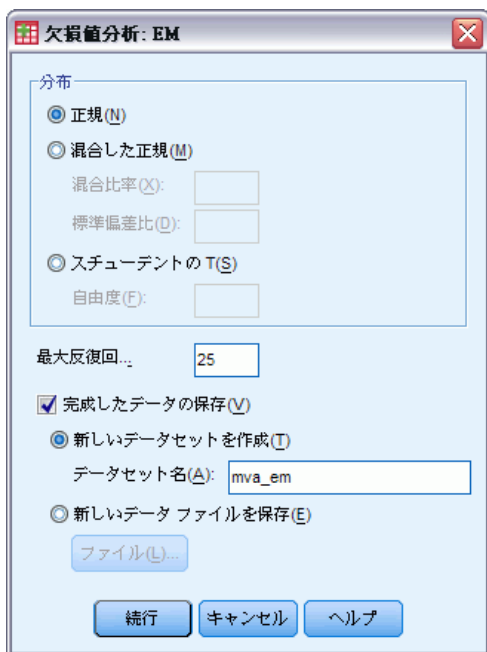
値が完全にランダムに欠損しているかどうか (MCAR) を検定するための Roderick J. A. Little のカイ 2 乗統計量が、EM 行列の脚注として印刷されます。この検定の帰無仮説は、データが完全にランダムに欠損しており、p 値は 0.05 レベルで有意となります。値が 0.05 未満の場合、データは完全にランダムには欠損していません。データはランダムに欠損している (MAR) か、ランダムに欠損していない (NMAR) のどちらかです。どちらか一方を仮定することはできないので、データを分析し、どのように欠損しているかを判断する必要があります。

### 回帰法 (因子分析)

この手法では、多重線型回帰推定値を計算します。また、ランダム成分で推定値を増加するオプションがあります。この手続きでは、各予測値に対し、ランダムに抽出した完全なケース、ランダム正規偏差、または t 分布の (残差平均平方の平方根で計測された) ランダム偏差からの残差を加えることができます。

## EM 推定のオプション

図 2-4  
[Missing Value Analysis EM (欠損値分析 - EM)] ダイアログ ボックス



EM 法では反復プロセスを使用し、欠損値のある量的（スケール）変数の平均、分散共分散行列、および相関を推定します。

**分布。** EM 法では、特定の分布の尤度に基づいて推定を行います。デフォルトでは、正規分布であると仮定します。分布の裾が正規分布より長いことがわかっている場合、手続きでスチューデントの  $t$  分布から自由度  $n$  で尤度関数が作成されるように要求できます。混合正規分布には、裾の長い分布も用意されています。混合正規分布の標準偏差の比と、2 つの分布の混合比率を指定します。混合正規分布では、分布の標準偏差のみが異なると仮定します。平均は同じであることが必要です。

**最大反復回数。** 最大反復回数を設定して、真の共分散を推定します。推定値が収束されなくても、この反復回数に達すると手続きは停止します。

**完了したデータの保存。** 代入された値を持つデータセットを欠損値の場所に保存することができます。ただし、代入値を使用する共分散に基づく統計量は、それぞれのパラメータ値を過小評価するので、注意が必要です。過小評価度は、同時に観測されていないケースの数に比例します。

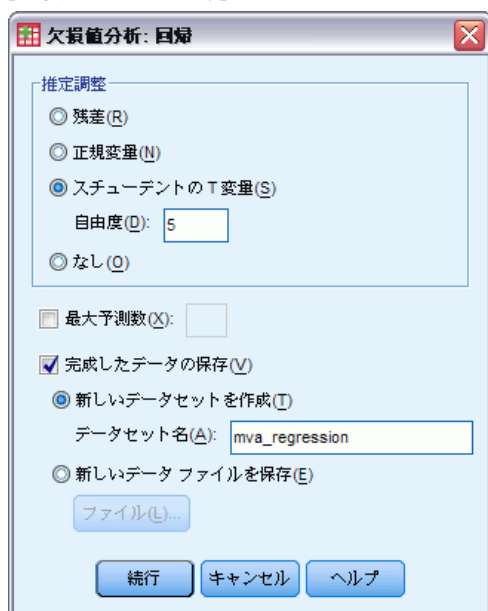
### EM オプションを指定するには

- ▶ [欠損値分析] メイン ダイアログ ボックスで、EM 法を使用して欠損値を推定表示する変数を選択します。

- ▶ [推定] グループから [EM] を選択します。
- ▶ 予測される変数と予測する変数を指定するには、[変数] をクリックします。詳細は、[p. 12 予測される変数と予測する変数](#) を参照してください。
- ▶ [EM] をクリックします。
- ▶ 必要な EM オプションを選択します。

## 回帰推定のオプション

図 2-5  
[欠損値分析 - 回帰] ダイアログ ボックス



回帰法では、多重線型回帰を使用して欠損値を推定します。予測された変数の平均、分散共分散行列、および相関行列が表示されます。

**推定の調整。** 回帰法により、回帰推定にランダムな成分を追加することができます。残差、正規変量、スチューデントの  $t$  変量、または調整なしを選択できます。

- **残差.** 誤差項を、完全なケースの観測された残差から無作為に選択して、回帰推定に加えます。
- **正規変量.** 期待値が 0 で、標準偏差が回帰の平均平方誤差項の平方根と等しい分布から、誤差項を無作為に引き出します。
- **スチューデントの T 変量.** 指定された自由度を有する  $T(n)$  分布から無作為に誤差項を引き出し、平均平方誤差の平方根 (RMSE) で尺度化します。

**最大予測数。** 推定過程で使用する予測 (独立) 変数の最大数を設定します。

**完了したデータの保存。** 回帰法による推定値で欠損値を置き換え、現在のセッションまたは外部の IBM® SPSS® Statistics データファイルにデータセットを書き込みます。

### 回帰オプションを指定するには

- ▶ [欠損値分析] メイン ダイアログ ボックスで、回帰法を使用して欠損値を推定表示する変数を選択します。
- ▶ [推定] グループから [回帰] を選択します。
- ▶ 予測される変数と予測する変数を指定するには、[変数] をクリックします。詳細は、 p.12 予測される変数と予測する変数 を参照してください。
- ▶ [回帰] をクリックします。
- ▶ 必要な回帰オプションを選択します。

## 予測される変数と予測する変数

図 2-6

[Missing Value Analysis Variables for EM and Regression (EM および回帰の欠損値分析変数)] ダイアログ ボックス



デフォルトでは、EM および回帰推定にすべての量的変数が使用されます。必要に応じて、推定で使用する予測される変数および予測する変数として、特定の変数を選択できます。特定の変数を両方のリストに含めることができますが、変数の使用を制限すべき状況もあります。たとえば、結果変数の値を推定することを控えたいと考える分析者もいます。また、異なる推定に異なる変数を使用し、手続きを複数回実行する場合があります。たとえば、看護師の評価と医師の評価に関するそれぞれの項目セットがある場合、看護師項目を使用してその欠損項目の推定を行い、もう一度別の推定を実行して医師項目の推定値を求めることができます。

回帰法を使用するときには、もう 1 つの懸念事項があります。多重回帰では、独立変数の大きなサブセットを使用すると、小さなサブセットより劣る予測値が得られる場合があります。したがって、変数を使用するには、投入される F 値制限として 4.0 を満たす必要があります。この制限は、シンタックスで変更できます。

### 予測される変数と予測変数を指定するには

- ▶ [欠損値分析] メイン ダイアログ ボックスで、回帰法を使用して欠損値を推定表示する変数を選択します。
- ▶ [推定] グループから [EM] または [回帰]を選択します。
- ▶ [変数] をクリックします。
- ▶ 予測される変数および予測変数として、すべての変数ではなく特定の変数を使用する場合は、[変数を選択] を選択して変数を該当するリストに移動します。

## MVA コマンドの追加機能

コマンド シンタックス言語を使用して、次のことも実行できます。

- MPATTERN、DPATTERN、または TPATTERN サブコマンドで DESCRIBE キーワードを使用して、欠損値パターン、データ パターン、および集計パターンに別々の記述変数を指定します。
- DPATTERN サブコマンドを使用して、データ パターン表に複数の並べ替え変数を指定します。
- DPATTERN サブコマンドを使用して、データ パターン表に複数の並べ替え変数を指定します。
- EM サブコマンドを使用して、許容度と収束基準を指定します。
- REGRESSION サブコマンドを使用して、許容度と投入される F 値を指定します。

- EM および REGRESSION サブコマンドを使用して、EM および回帰に異なる変数リストを指定します。
- TTESTS、TABULATE、および MISMATCH のそれぞれに、表示されるケースを非表示にするパーセントを指定します。

複雑なシンタックス情報については、「コマンド シンタックス リファレンス」を参照してください。

# 多重代入












多重代入の目的は欠損値の代わりに使用可能な値を生成して、データの“完全”なセットをいくつか作成することです。多重代入データセットで使用される分析手続きは、それぞれの“完全”なデータセットに対する出力に加え、元のデータセットに欠損値がなかった場合の結果を推定するプールされた出力を作成します。通常、これらのプールされた結果は、単一代入方法により求められる結果よりも正確です。

**分析変数。** 分析変数は次のようになります。

- **名義データ。** 値がランキングなどを持たないカテゴリを表しているとき、名義（変数）として取り扱うことができます。たとえば、従業員の会社の所属などです。名義変数の例としては、地域やジップコードや所属宗教などがあります。
- **順序データ。** 値がランキングをもったカテゴリを表しているとき、変数を順序として取り扱うことができます。たとえば、「かなり不満」から「かなり満足」までのようなサービス満足度のレベルなどです。順序変数の例としては、満足度や信頼度を表す得点や嗜好得点などです。
- **スケールデータ。** 値が有意な基準を持った順序カテゴリを表しているとき、変数をスケール（連続型）として扱うことができます。値間の距離の比較などに適切です。スケール変数の例としては、年齢や、千ドル単位で表した所得があります。

この手続きは適切な尺度がすべての変数に割り当てられると仮定します。ただし、ソース変数リスト内の変数を右クリックしコンテキストメニューから尺度を選択して、変数の尺度を一時的に変更することができます。

変数リストで各変数の隣にあるアイコンは、次のような尺度とデータ型を表します。

	数値	String	Date	Time
スケール（連続）		利用不可		
順序				
名義				

**度数による重み付け。** 度数（複製）による重み付けは、この手続きによって処理されます。負の値または 0 の複製による重み付けを持つケースは無視されます。非整数値の重み付けは最も近い整数に丸められます。

**分析の重み付け:** 分析（回帰または抽出）の重み付けは、欠損値の要約および代入モデルの当てはめに組み込まれています。負の値または 0 の分析の重み付けを持つケースは除外されます。

**コンプレックス サンプル** 多重代入の手続きは、分析重み変数の形式の最終抽出重みを受け入れることができますが、明確にはストラータ、クラスまたはコンプレックス サンプリング構造を処理しません。コンプレックス サンプリング手続きは多重代入データセットを自動的には分析しないことに注意してください。プールのサポートのリストの詳細は、「[多重代入データの分析](#) p.31 .」を参照してください。

**欠損値。** ユーザ欠損値およびシステム欠損値は両方とも無効な値として取り扱われます。つまり、両タイプの欠損値は値が代入される際に置き換えられ、両方とも代入モデルで予測変数として使用される変数の無効な値として取り扱われます。欠損値の分析では、ユーザー欠損値とシステム欠損値は同様に欠損として取り扱われます。

**結果の再現（欠損データ値の代入）。** 代入の結果を正確に再現するには、同じ手続きの設定を行うだけでなく、乱数ジェネレータに同じ初期化の値、同じデータの順序、同じ変数の順序を使用します。

- **乱数ジェネレータ。** この手続きでは、代入値の計算時に乱数ジェネレータを使用します。今後同じランダム化された結果を再生成するには、[欠損データ値を代入] 手続きをそれぞれ実行する前に乱数ジェネレータに同じ初期化の値を使用します。
- **ケースの並び順。** 値はケースの並び順に従って代入されます。
- **変数の順序。** 完全条件指定（FCS）代入方法は、[分析変数] リストで指定された順序で値を代入します。

多重代入専用のダイアログが 2 つあります。

- **[パターン分析]** は、データ中の欠損値パターンの記述方法を提供します。これは、代入前の検討段階で有用です。
- **[欠損データ値を代入]** は、多重代入の生成に使用されます。多重代入データセットをサポートする手順を使用して、完全なデータセットを分析することができます。多重代入データセットの分析およびこれらのデータをサポートする手順のリストの詳細については、[多重代入データの分析](#) p.31 を参照してください。



## パターン分析

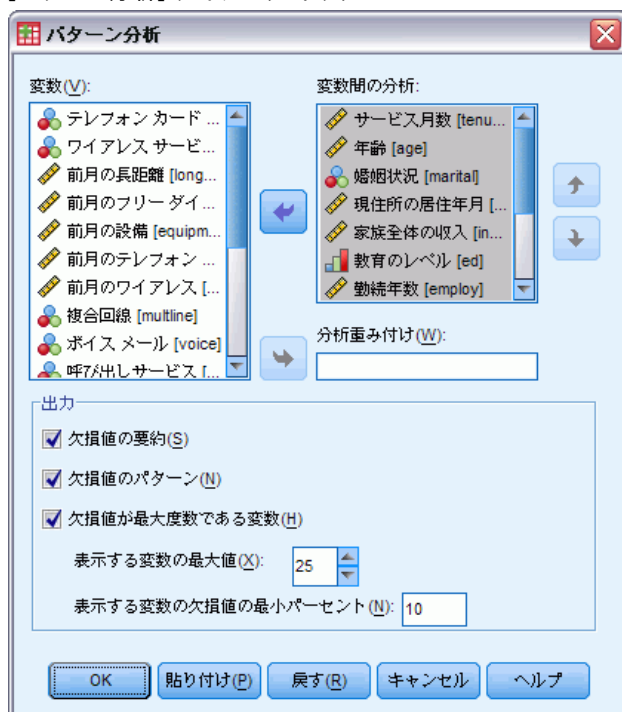
[パターン分析] は、データ中の欠損値パターンの記述方法を提供します。これは、代入前の検討段階で有用です。

**例:**あるデータ通信プロバイダは、顧客データベースにあるサービス利用パターンをさらに良く理解したいと考えています。彼らは、顧客が使用するサービスの完全なデータセットを持っていますが、この企業が収集した人口統計情報には、多くの欠損値が存在します。欠損値のパターンの分析は代入の次のステップを決定するのに役立ちます。詳細は、5章 p.55 [多重代入を使用したデータセットの比較と分析](#) を参照してください。

### 欠損データのパターンを分析するには

メニューから次の項目を選択します。  
分析(A) > 多重代入 > パターン分析...

図 3-1  
[パターン分析] ダイアログ ボックス



- ▶ 2 つ以上の分析変数を選択します。この手順により、これらの変数の欠損データのパターンが分析されます。

## オプション設定

**分析の重み付け:** この変数には、分析（回帰または抽出）の重み付けが含まれます。この手順では、欠損値の要約に分析の重み付けが含まれます。負の値または 0 の分析の重み付けを持つケースは除外されます。

**出力。** オプションの出力は次のとおりです。

- **欠損値の要約。** ここでは、パネルに分けられた円グラフが表示されます。このグラフでは、分析変数、ケース、または 1 つ以上の欠損値を持つ個別のデータ値の数およびパーセントが示されます。
- **欠損値のパターン。** ここでは、作表された欠損値のパターンが表示されます。それぞれのパターンは、分析変数に不完全および完全なデータの同じパターンを持つケースのグループに対応します。この出力を使用して、データに対して単調な代入を使用することができるか判定することができます。使用できない場合は、データが単調なパターンにどのくらい近似しているのか判断できます。この手順は分析変数に順序を付けて、単調なパターンを明らかにしたり、単調なパターンに近づけたりします。並び替えの後に非単調パターンが存在しない場合は、分析変数が順序付けられた際にデータは単調なパターンを持っていると結論付けることができます。
- **欠損値の最高度数を持つ変数。** ここでは、欠損値のパーセントの大きいもの順に並び替えられた分析変数の表が表示されます。この表には、スケール変数の記述統計（平均値および標準偏差）が含まれます。

表示する変数の最大数および表示に含める変数の欠損の最小パーセントを制御することができます。両方の基準を満たす変数のセットが表示されます。たとえば、変数の最大数を 50 に、最小の欠損パーセントを 25 に設定すると、最大で 50 個の 25% 以上の欠損値を持つ変数を含む表の表示が要求されます。60 個の分析変数があり、そのうちの 15 個のみが 25% 以上の欠損値を持っている場合、出力には 15 個の変数のみが含まれます。

## 欠損データ値の代入

[欠損データ値を代入] は、多重代入の生成に使用されます。多重代入データセットをサポートする手順を使用して、完全なデータセットを分析することができます。多重代入データセットの分析およびこれらのデータをサポートする手順のリストの詳細については、[多重代入データの分析](#) p. 31 を参照してください。

**例:**あるデータ通信プロバイダは、顧客データベースにあるサービス利用パターンをさらに良く理解したいと考えています。彼らは、顧客が使用するサービスの完全なデータセットを持っていますが、この企業が収集した人口統計情報には、多くの欠損値が存在します。さらに、この値は完全にランダムで欠損しているわけではないため、多重代入を利用して、デー

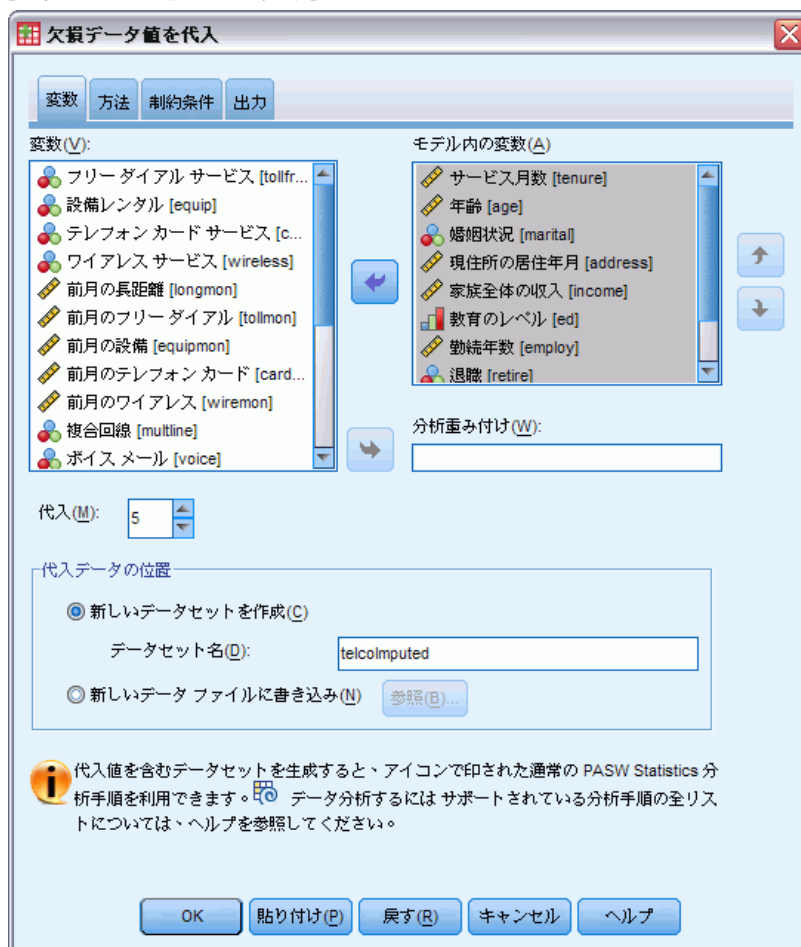
タセットを完成させます。詳細は、5章 p.55 多重代入を使用したデータセットの比較と分析 を参照してください。

### 欠損データ値を代入するには

メニューから次の項目を選択します。

分析(A) > 多重代入 > 欠損データ値を代入...

図 3-2  
[欠損データ値を代入 - 変数] タブ



- ▶ 2 つ以上の変数を代入モデルで選択します。この手順により、これらの変数の欠損データに複数の値が代入されます。
- ▶ 計算する代入の数を指定します。この値のデフォルトは 5 です。
- ▶ 代入されたデータを書き込むデータセットまたは IBM® SPSS® Statistics フォーマット データ ファイルを指定します。

出力データセット欠損値を持つオリジナルのケース データと各代入に対する代入値を持つケースのセットから成り立っています。例えば、オリジナル データセットが 100 のケースがあり、5 の代入がある場合、出力データセットは 600 ケースになります。入力データセットのすべての変数は、出力データセットに含まれます。既存の変数の辞書のプロパティ（名前、ラベルなど）は新しいデータセットにコピーされます。ファイルは、新しい数値変数で代入を示す（元のデータは 0、代入値を持つケースは 1..n）Imputation\_を含みます。

この手順は、出力データセットが作成される際に自動的にImputation\_変数を分割変数として定義します。この手順の実行の際に分割が有効になっている場合、出力データセットには、分割変数値の組み合わせに対する代入がそれぞれ 1 セット含まれます。

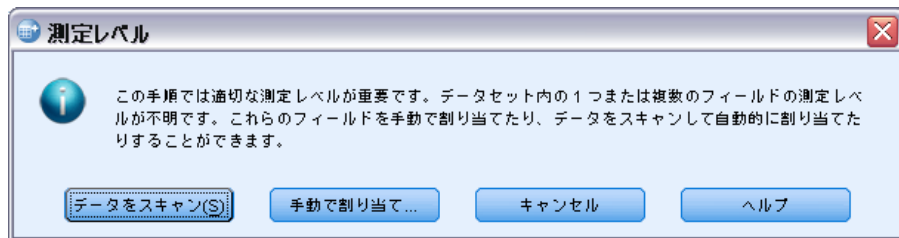
### オプション設定

**分析の重み付け:** この変数には、分析（回帰または抽出）の重み付けが含まれます。この手順には、回帰における分析の重み付けおよび欠損値の代入に使用される分類モデルが含まれます。分析の重み付けは、平均値、標準偏差、標準誤差などの代入値の要約でも使用されます。負の値または 0 の分析の重み付けを持つケースは除外されます。

### 測定レベルが不明なフィールドです。

データセットの 1 つまたは複数の変数（フィールド）の尺度が不明な場合、尺度の警告が表示されます。尺度はこの手順の結果の計算に影響を与えるため、すべての変数に尺度を定義する必要があります。

図 3-3  
尺度の警告

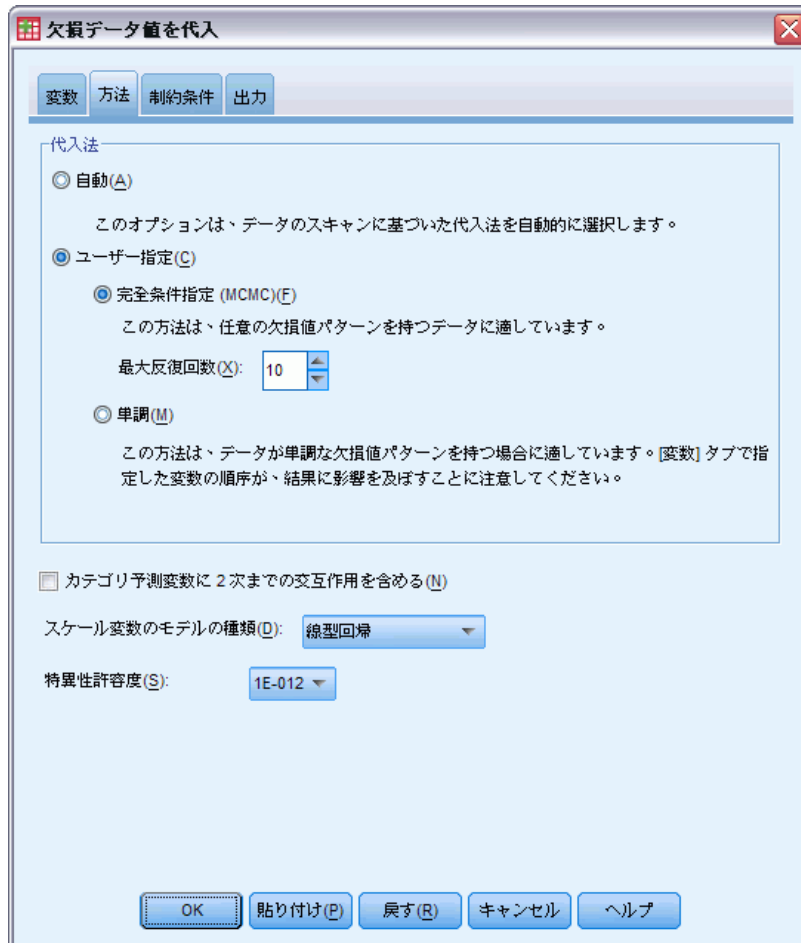


- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの [変数ビュー] でも、尺度を割り当てることができます。

尺度がこの手順で重要であるため、すべてのフィールドに尺度が定義されるまで、ダイアログにアクセスしてこの手順を実行することはできません。

## 方法

図 3-4  
[欠損データ値を代入 - 方法] タブ



[方法] どのようにタブは使用されたモデルの型を含む欠損値が代入されるかを指定します。カテゴリ予測変数はダミー コード化された指示変数です。

**代入方法。**自動方法はデータをスキャンして、データに単調なパターンの欠損値が見られる場合は単調方法を使用します。それ以外の場合は、完全条件指定が使用されます。使用する方法が決まっている場合、その方法をユーザー指定の方法として指定することができます。

- **完全条件指定。**これは、欠損データのパターンが任意（単調または非単調）の場合に使用可能な反復マルコフ連鎖モンテカルロ（MCMC）方法です。

変数リストで指定された順序の各変数と反復では、完全条件指定（FCS）方法が、予測変数としてモデル内のすべての他の使用可能な変数を使用する 1 変数（単一従属変数）に当てはまり、当てはまった変数の欠損値が代入されます。この方法は最大反復数に達するまで続き、最大反復数での代入値が代入データセットに保存されます。を参照してください。

**最大反復回数。**これにより、FCS 方法で使用されるマルコフ連鎖により取得される反復または“ステップ”の数が指定されます。FCS 方法が自動的に選択されている場合、デフォルトの反復回数である 10 が使用されます。FCS を明示的に選択する場合、反復回数を指定することができます。マルコフ連鎖が収束していない場合、反復回数を増やす必要があります。[出力] タブでは、FCS 反復履歴データを保存して、収束を評価するためにプロットすることができます。

- **単調。**これはデータに単調なパターンの欠損値がある場合のみ利用できる非反復方法です。変数に非欠損値があり、先行するすべての変数にも非欠損値がある変数を並び替えられる場合、単調なパターンが存在します。これを[ユーザー指定]として指定するときは、単調なパターンを表示する順序のリストの変数を必ず否定してください。

単調順序の各変数では、単調方法が、予測変数としてモデル内の先行する変数を使用する 1 変数（単一従属変数）に当てはまり、当てはまった変数の欠損値が代入されます。これらの代入された値は代入データセットに保存されます。を参照してください。

**2 次の交互作用を含める。**代入方法が自動的に選択される場合、各変数の代入モデルには予測変数に対する定数項や主効果が含まれます。特定の方法を選択する場合、オプションでカテゴリ予測変数の中から 2 次までの交互作用を含めることができます。

**スケール変数のモデルタイプ。**代入方法が自動的に選択される場合、線形回帰がスケール変数の 1 変数モデルとして使用されます。特定の方法を選択する場合は、代わりに予測平均値一致（PMM）をスケール変数のモデルとして選択することができます。PMM は、最も近い観測値への回帰モデルを使用して計算される代入値を一致させる線形回帰の可変配列です。

ロジスティック回帰は、常にカテゴリ変数に対する 1 変数のモデルとして使用されます。モデルの種類にかかわらず、カテゴリ予測変数はダミーコード化された指示変数を使用して扱われます。

**特異性許容度。**特異（または不可逆）行列は、2 つ以上の列が線型従属になっており、推定アルゴリズムにとって大きな不都合となることがあります。準特異行列についても、十分な結果が得られない場合があります。そのためこの手続きでは、行列式が許容値未満である行列は、特異行列とみなされます。正の値を指定します。

## 制約

図 3-5  
[欠損データ値を代入 - 制約] タブ

The screenshot shows the '欠損データ値を代入' dialog box with the '制約条件' tab selected. The '変数の要約のデータのスキャン' section includes a 'データをスキャン(S)' button, a checkbox for 'スキャンするケースの数を制限する(L)', and a 'ケース(C): 5000' field. Below is a table for '変数の要約(V):' with columns for 'モデル内の変数', '欠損パーセント', '最小観測値', and '最大観測値'. The '制約条件を定義(D):' section contains a table with columns for 'モデル内の変数', '役割', '最小値', '最大値', and '丸め'. Below this are checkboxes for '大量の欠損データを持つ変数を除外(E)', and input fields for '欠損の最大パーセント(M)', 'ケースの最大抽出数(A)', and 'パラメータの最大抽出数(X)'. An information icon and a note are also present.

モデル内の変数	欠損パーセント	最小観測値	最大観測値
tenure			
age			
marital			
address			

スキャンされたケース:none

モデル内の変数	役割	最小値	最大値	丸め
tenure	予測変数として代入および...			
gender	予測変数として代入および...			
reside	予測変数として代入および...	1		1
lninc	予測変数として代入および...	0		

大量の欠損データを持つ変数を除外(E)  
欠損の最大パーセント(M):   
ケースの最大抽出数(A):   
パラメータの最大抽出数(X):

パラメータの最大抽出数を増やすと、分析時間が有意に長くなります。

Buttons: OK, 貼り付け(P), 戻す(R), キャンセル, ヘルプ

[制約条件] タブを指定すると、代入中の変数の役割を制限することができます。また、スケール変数の代入値を妥当な範囲に制限することもできます。加えて、分析を欠損値の最大パーセントより少ない変数に制限することもできます。

**変数要約のデータのスキャン。**[データのスキャン]をクリックすると、このリストには、分析変数とそれぞれの分析変数の観測された欠損パーセント、最小値、および最大値が表示されます。[ケース] テキスト ボックスでの指定に応じて、要約はすべてのケースに基づくか、あるいは最初の n ケースのスキャンに限定されます。[データを再スキャン] をクリックすると、分布の要約が更新されます。

### 制約の定義

- **役割。** これを使用してカスタマイズすると、変数のセットが代入されるようにしたり、予測変数として扱われるようにしたりすることができます。通常、各分析変数は、代入モデルの予測変数および独立変数の両方であると考えられます。役割を使用して、予測変数としてのみ使用したい変数への代入を無効にしたり、予測変数として使用されないように変数を除外（代入のみ）して、予測モデルをより小さくしたりすることができます。これは、カテゴリ変数や予測変数として使用される変数に指定することのできる唯一の制約です。
- **最小と最大。** これらの列では、スケール変数への最大または最小の代入値を指定することができます。代入値が指定された範囲から外れた場合、範囲に含まれる値が得られるかあるいは最大引き出し数に達する（下記の最大引き出し数を参照）まで、この手順により別の値が引き出されます。これらの列は、線型回帰がスケール変数モデルの種類として [方法] タブで選択されている場合のみ使用できます。
- **丸め。** いくつかの変数はスケールとして使用されますが、自然に制限される変数を持っています。例えば、世帯の人数は整数ですが、コンビニなどで使用する金額は小数点以下はありません。この列により許容できる最小の単位を指定できます。例えば、整数値を得るために、丸め単位として 1 を指定し、近似値の銭に丸める値を得るには、0.01 を指定します。一般値は、丸められた単位の最も近い整数の倍数に丸められます。以下の表には、6.64823（丸め前）に対して異なる丸め値がどのように作用するか示されています。

丸め単位	6.64832 が丸められる数値
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

**大量の欠損データを持つ変数を除外する。** 分析変数が代入モデルを推定するために十分なデータを持っている場合、通常、これらは所持している欠損値の数に関わらず予測変数として代入され、使用されます。高いパーセントの欠損値を持っている変数を除外するように指定することができます。たとえば、50 を [欠損の最大パーセント] として指定した場合、50% より大きい



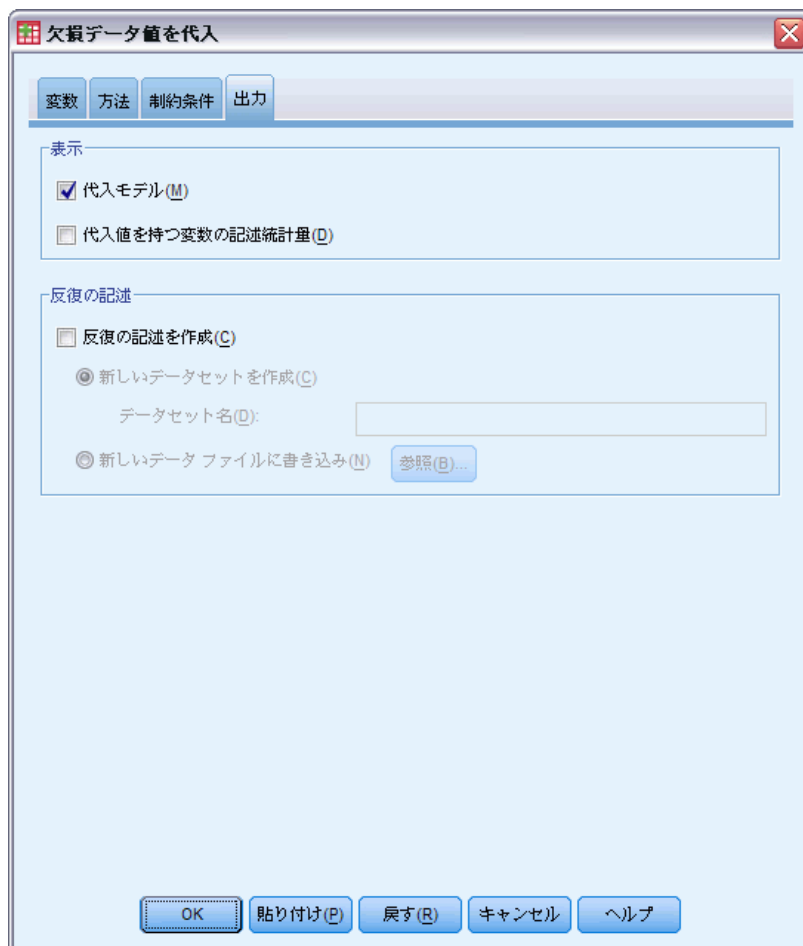
欠損値を持っている分析変数は代入されず、代入モデルで予測変数として使用されることもありません。

**最大引き出し数。** スケール変数の代入値に最小または最大値が指定されている場合（上記の**最大と最小**を参照）は、指定された範囲に収まる値のセットが見つかるまで、この手順によりケースに対して値が引き出されます。ケースごとに指定された引き出し数内で値のセットが得られない場合、この手順により別のセットのモデルパラメータが引き出され、ケース引き出しプロセスが繰り返されます。指定されたケースおよびパラメータ引き出し回数内で範囲内の値のセットが得られない場合は、エラーが発生します。

これらの値を増やすと、処理時間も長くなることに注意してください。この手順の実行に長い時間が掛かったり、適切な引き出しを探し出せない場合は、指定された最小値と最大値が適切かどうか確認してください。

## 出力

図 3-6  
[欠損データ値を代入 - 出力] タブ



**表示。** 出力表示を制御代入指定に関連する表、反復（完全条件指定方法の場合）、代入された従属変数、代入から除外された従属変数、および代入順序を含む、全体的な代入要約は常に表示されます。分析変数の制約が指定されている場合は、これも同様に表示されます。

- **代入モデル。** ここでは、従属変数と予測変数の代入モデルに加え、1 変量モデル タイプ、モデル効果、および代入値の数が表示されます。
- **記述統計。** ここでは、値が代入される従属変数の記述統計が表示されます。スケール変数の場合、記述統計には、（代入前の）元の入力データの平均値、度数、標準偏差、最小値、および最大値に加え、（代入による）代入値と完全なデータ（元の値と代入値の両方 - 代入による）が含まれます。カテゴリ変数の場合、記述統計には、（代入前の）元の入力

データのカテゴリ別の度数とパーセントに加え、(代入による) 代入値と完全なデータ (元の値と代入値の両方 - 代入による) が含まれます。

**反復履歴。** 完全条件指定代入方法が使用された場合は、FCS 代入の反復履歴データを含むデータセットを要求することができます。このデータセットには、反復による平均値と標準偏差、および値が代入される各スケール従属変数への代入が含まれます。データをプロットして、モデル収束の評価に利用することができます。詳細は、5 章 p.76 FCS 収束の確認を参照してください。

## MULTIPLE IMPUTATION コマンドの追加機能

コマンド シンタックス言語を使用して、次のことも実行できます。

- 記述統計が表示される変数のサブセットを指定します (IMPUTATIONSUMMARIES サブコマンド)。
- 欠損パターンの分析と手順の 1 回の実行における代入を両方指定します。
- モデル パラメータの最大数を指定するとその変数も代入できます (MAXMODELPARAM キーワード)。

複雑なシンタックス情報については、「コマンド シンタックス リファレンス」を参照してください。

## 多重代入データを使用した作業

多重代入 (MI) データセットが作成されると、変数ラベル付きの Imputation\_、代入番号 (Imputation\_) が追加され、データセットはこの変数により昇順に並び替えられます。元のデータセットのケースは値が 0 です。代入された値のケースには 1 ~ M の番号が付きます。M は代入の数を示します。

データセットを開いたときに 代入 (Imputation\_) がある場合、これはデータセットが MI データセットである可能性を示しています。

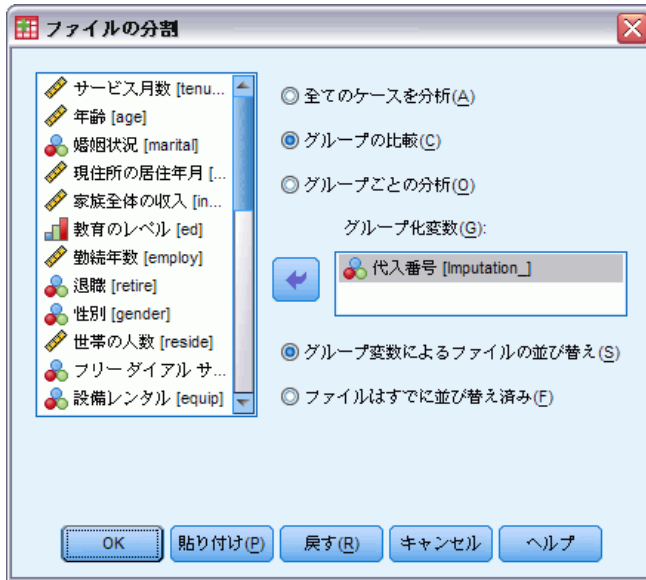
### 分析用の多重代入データセットのアクティブ化

分析で MI データセットして取り扱われるために、データセットは、[グループの比較] オプションで 代入 (Imputation\_) をグループ変数として使用して、分割されている必要があります。他の変数の分割も定義することができます。

メニューから次の項目を選択します。

データ > ファイルの分割(F)...

図 3-7  
[ファイルの分割] ダイアログ ボックス



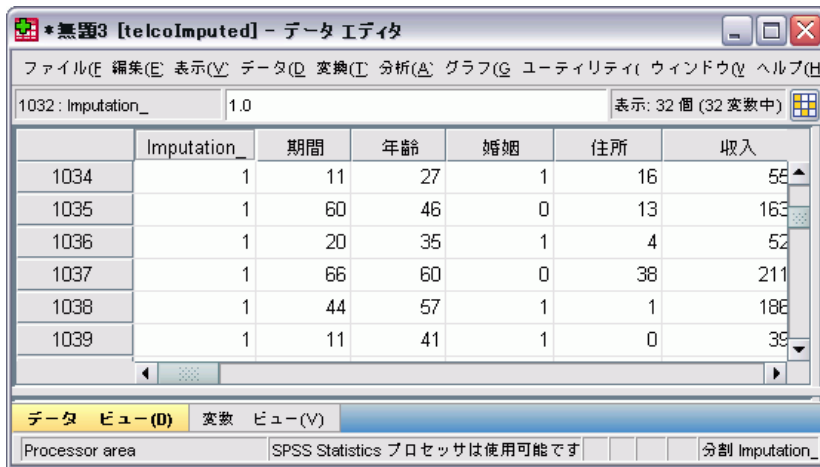
- ▶ [グループの比較] を選択します。
- ▶ [代入番号 (Imputation\_)] をケースをグループ化する変数として選択します。

または、マーキングを有効にする際（下記を参照）に、積によりファイルが [代入番号 (Imputation\_)] で自動的に分割されます。

### 観察値からの代入値の区別

セルの背景色、フォント、太字などを代入値に使用して、代入値を観察値から区別することができます。有効なマーキングについての詳細は、[多重代入オプション](#) p. 38 を参照してください。代入欠損値のある現在のセッションの新しいデータセットを作成するときは、デフォルトではマーキングは有効にされています。デフォルトでは、代入を含むデータ ファイルを開くときは、マーキングは無効にされています。

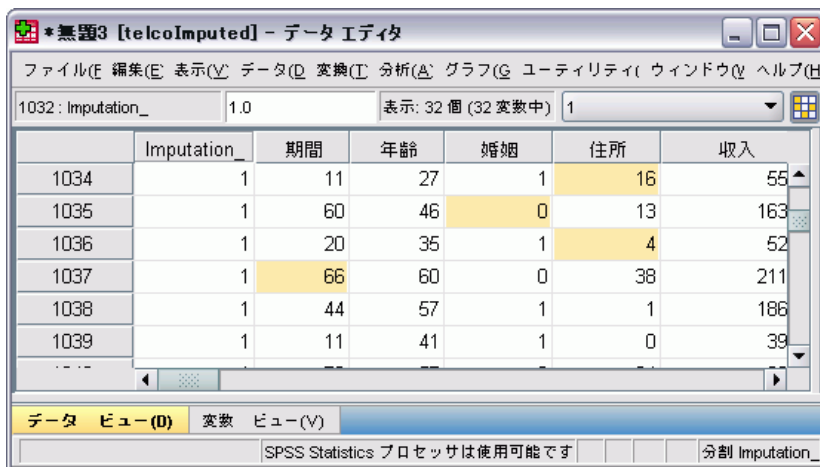
図 3-8  
代入マーキングが無効なデータ エディタ



マーキングを有効にするには、データ エディタのメニューから次の項目を選択します。

表示 > [代入データをマーク]

図 3-9  
代入マーキングが有効なデータ エディタ



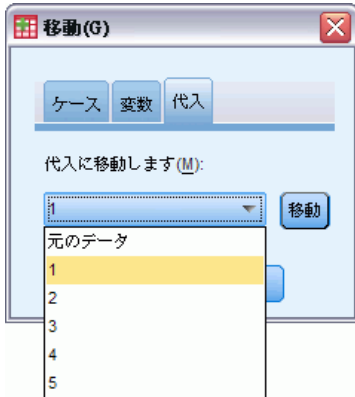
また、データ エディタのデータビュー内で編集バーの右端にある代入マーキング ボタンをクリックして、マーキングを有効にすることもできます。

### 代入間の移動

- ▶ メニューから次の項目を選択します。  
編集 > 代入に移動(M)...

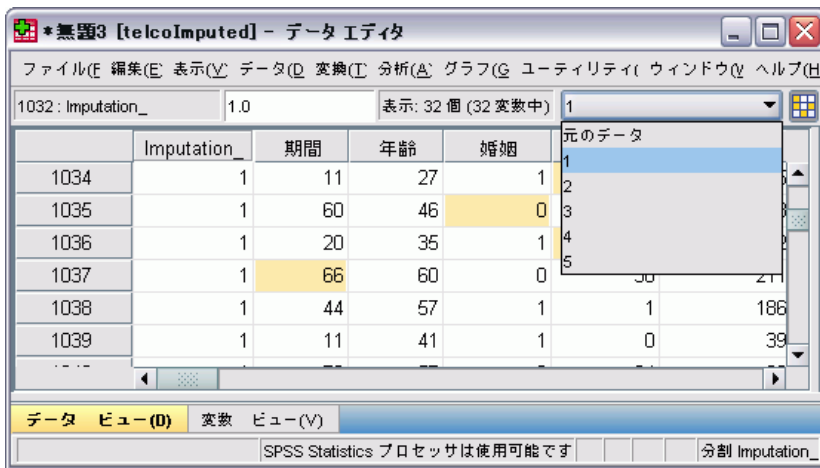
- ▶ ドロップダウン リストから代入（またはオリジナルデータ）を選びます。

図 3-10  
ダイアログ ボックスに移動



または、データ エディタのデータ ビューにある編集バーのドロップダウン リストから代入を選択することもできます。

図 3-11  
代入マーキングが有効なデータ エディタ



代入を選択する際、相対的なケース位置は保持されます。たとえば、オリジナルのデータセットに 1000 ケースがある場合、最初の代入の 34 番目のケースであるケース 1034 がグリッドの一番上に表示されています。代入 2 をドロップダウンで選択すると、2 番目の代入の 34 番目のケースであるケース 2034 がグリッドの一番上に表示されます。元のデータをドロップダウンで選択すると、ケース 34 がグリッドの一番上に表示されます。また、代入を切り替える際はカラムの位置も保持されるので、代入ごとの値を容易に比較することができます。

### 代入値の変換と編集

代入データの変換を行う必要があることがあります。たとえば、給与変数のすべての値のログを取得して、結果を新しい変数に保存したい場合などです。代入されたデータを使用して計算された値は、元のデータを使用して計算した値と違う場合、代入値として扱われます。

データ エディタのセルで代入値を編集した場合、そのセルは引き続き代入済みとして処理されます。このようにして代入値を編集することは推奨されていません。

## 多重代入データの分析

多重代入データセットの分析から得られた結果のプールは、多くの手順によりサポートされています。代入マーキングが有効な場合、プールをサポートする手順の隣に特殊なアイコンが表示されます。たとえば、[分析] メニューの [記述統計] サブメニューにある [度数分布表]、[記述統計量]、[探索的]、および [クロス集計表] はすべてプールをサポートしますが、[比データ]、[正規 P-P プロット]、および [正規 Q-Q プロット] はサポートしません。

図 3-12  
代入マーキングが有効な分析メニュー



表の出力とモデル PMML の両方をプールすることができます。プールされた出力を要求するための新しい手順はありません。多重代入の出力の包括的な制御は、[オプション] ダイアログの新しいタブで行えます。

- **表出力のプール** 多重代入 (MI) データセットでサポートされている手順を実行する場合、デフォルトで、結果はそれぞれの代入、元の（代入されていない）データ、および代入での変数が考慮されたプールされている（最終）結果に対して自動的に生成されます。プールされた統計は手順により変化します。
- **PMML のプール**。プールされた PMML を PMML をエクスポートするサポートされている手順から取得することもできます。プールされた PMML は同じ方法で要求され、プールされていない PMML の代わりに保存されます。

サポートされていない手順は、プールされた出力もプールされていない PMML ファイルのどちらも作成しません。



## プールのレベル

出力は 2 つあるレベルの片方を使用してプールされます。

- **Naïve 組み合わせ。**プールされたパラメータのみが利用可能です。
- **1 変量の組み合わせ。**プールされたパラメータ、その標準誤差、検定統計量と有効自由度、p 値、信頼区間、およびプール診断（欠損情報の割合、相対効率、分散の相対増加）が利用可能な場合表示されます。

係数（回帰と相関）、平均値（平均値の差）、度数はプールされます。統計の標準誤差が利用できる場合は、1 変量プールが使用され、他の場合は、naïve プールが使用されます。

## プールをサポートする手順

次の手順では、MI データセットがそれぞれの出力に指定されたプールのレベルでサポートされています。

### 度数

- 統計量テーブルでは、1 変量プールにおける平均値（標準誤差の平均値も要求されている場合）と Naïve プールにおける有効 N と欠損 N がサポートされています。
- 度数分布表では、Naïve プールにおける度数がサポートされています。

### 記述統計

- 記述統計量テーブルでは、1 変量プールにおける平均値（標準誤差の平均値も要求されている場合）と Naïve プールにおける N がサポートされています。

### クロス集計表

- クロス表テーブルでは、Naïve プールにおける度数がサポートされています。

### 平均値

- レポートテーブルでは、1 変量プールにおける平均値（標準誤差の平均値も要求されている場合）と Naïve プールにおける N がサポートされています。

### 1 サンプルの t 検定

- 統計量テーブルでは、1 変量プールにおける平均値と Naïve プールにおける N がサポートされています。
- 検定テーブルでは、1 変量プールにおける平均値の差がサポートされます。

### 独立したサンプルの t 検定

- グループ統計量テーブルでは、1 変量プールにおける平均値と Naïve プールにおける N がサポートされています。
- 検定テーブルでは、1 変量プールにおける平均値の差がサポートされます。

### 対応のあるサンプルの t 検定

- 統計量テーブルでは、1 変量プールにおける平均値と Naïve プールにおける N がサポートされています。
- 相関テーブルでは、Naïve プールにおける相関と N がサポートされています。
- 検定テーブルでは、1 変量プールにおける平均値がサポートされます。

### 一元配置分散分析

- 記述統計量テーブルでは、1 変量プールにおける平均値と Naïve プールにおける N がサポートされています。
- 対比検定テーブルでは、1 変量プールにおける対比の値がサポートされています。

### 線型混合モデル

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。
- 固定効果の推定値テーブルでは、1 変量プールにおける推定値がサポートされます。
- 共分散パラメータの推定値テーブルでは、1 変量プールにおける推定値がサポートされます。
- [推定周辺平均: 推定] テーブルでは、1 変量プールにおける平均値がサポートされます。
- [推定周辺平均: ペアごとの比較] テーブルでは、1 変量プールにおける平均値の差がサポートされます。

**一般化線型モデルと一般化推定方程式。**これらの手順ではプールされた PMML がサポートされます。

- カテゴリ変数情報テーブルでは、Naïve プールにおける N とパーセントがサポートされています。
- 連続変数情報テーブルでは、Naïve プールにおける N と平均値がサポートされています。
- パラメータ推定値テーブルでは、1 変量プールにおける係数と B がサポートされています。
- [推定周辺平均: 推定係数] テーブルでは、Naïve プールにおける平均値がサポートされます。

- [推定周辺平均: 推定] テーブルでは、1 変量プールにおける平均値がサポートされます。
- [推定周辺平均: ペアごとの比較] テーブルでは、1 変量プールにおける平均値の差がサポートされます。

## 2 変量の相関分析

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。
- 相関テーブルでは、1 変量プールにおける相関と N がサポートされています。プーリングの前に相関は Fisher の z 変換を使用して変換し、プーリングの後、逆方向に変換します。

## 偏相関分析

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。
- 相関テーブルでは、Naïve プールにおける相関がサポートされています。

**線型回帰。**この手順ではプールされた PMML がサポートされます。

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。
- 相関テーブルでは、Naïve プールにおける相関と N がサポートされています。
- 係数テーブルでは、1 変量プールにおける B と Naïve プールにおける相関がサポートされています。
- 相関係数テーブルでは、Naïve プールにおける相関がサポートされています。
- 残差統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。

**2 項ロジスティック回帰。**この手順ではプールされた PMML がサポートされます。

- 式に含まれる変数テーブルでは、1 変量プールにおける B がサポートされています。

**多項ロジスティック回帰。**この手順ではプールされた PMML がサポートされます。

- パラメータ推定値テーブルでは、1 変量プールにおける係数と B がサポートされています。

### 順序回帰

- パラメータ推定値テーブルでは、1 変量プールにおける係数と B がサポートされています。

**判別分析。**この手順ではプールされたモデルの XML がサポートされます。

- グループ統計量テーブルでは、Naïve プールにおける平均値と有効な N がサポートされています。
- プールされたグループ内行列テーブルでは、Naïve プールにおける相関がサポートされています。
- 正準判別関数係数テーブルでは、Naïve プールにおける標準化されていない係数がサポートされています。
- グループの重心における関数テーブルでは、Naïve プールにおける標準化されていない係数がサポートされています。
- 分類関数係数テーブルでは、Naïve プールにおける係数がサポートされています。

### カイ 2 乗検定

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。
- 度数分布表では、Naïve プールで観測された N がサポートされています。

### 2 項検定

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。
- 検定テーブルでは、Naïve プールにおける N、観測比率、および検定比率がサポートされています。

### ラン検定

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。

### 1 サンプルによる Kolmogorov-Smirnov 検定

- 記述統計量テーブルでは、Naïve プールにおける平均値と N がサポートされています。

### 2 個の独立サンプルの検定

- ランクテーブルでは、Naïve プールにおける平均ランクと N がサポートされています。
- 度数分布表では、Naïve プールにおける N がサポートされています。

### 複数の独立サンプルの検定

- ランクテーブルでは、Naïve プールにおける平均ランクと  $N$  がサポートされています。
- 度数分布表では、Naïve プールにおける度数がサポートされています。

## 2 個の対応サンプルの検定

- ランクテーブルでは、Naïve プールにおける平均ランクと  $N$  がサポートされています。
- 度数分布表では、Naïve プールにおける  $N$  がサポートされています。

## 複数の対応サンプルの検定

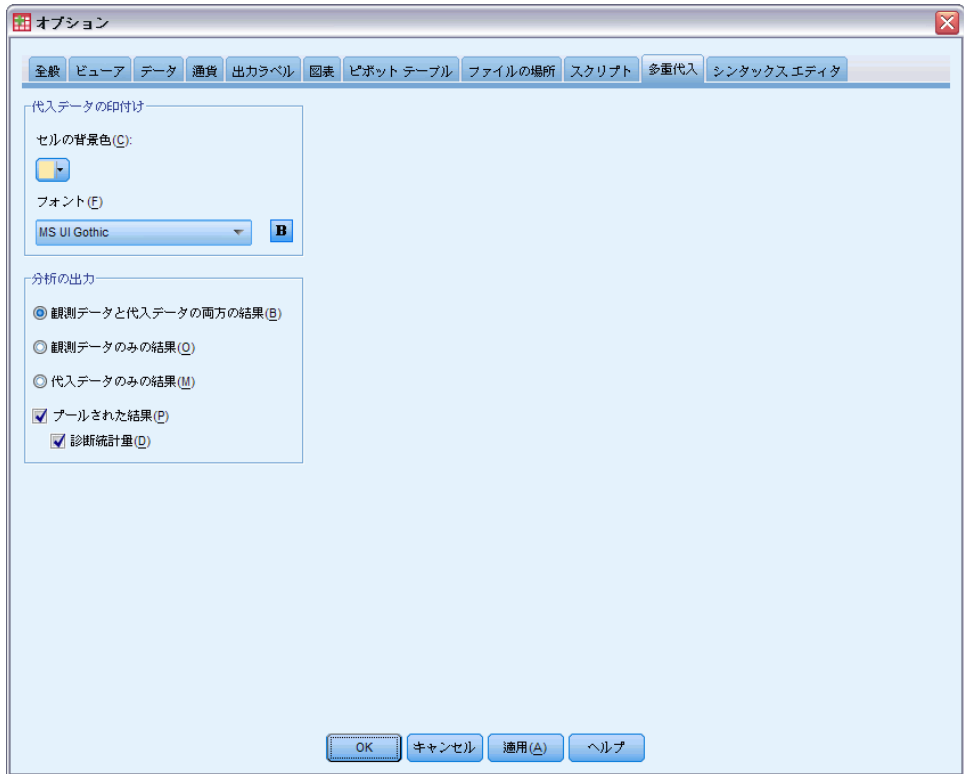
- ランクテーブルでは、Naïve プールにおける平均ランクがサポートされています。

**Cox 回帰分析。**この手順ではプールされた PMML がサポートされます。

- 式に含まれる変数テーブルでは、1 変数プールにおける  $B$  がサポートされています。
- 共変量平均値テーブルでは、Naïve プールにおける平均値がサポートされています。

## 多重代入オプション

図 3-13  
[オプション] ダイアログ ボックス: [多重代入] タブ



[多重代入] タブでは、多重代入に関連する 2 種類の設定を制御します。

**代入データの外觀。**デフォルトで、代入されたデータを含むセルは、代入されていないデータを含むセルとは異なる背景色を持っています。代入されたデータは特徴的な外觀を持っているので、それらのセルはデータセットをスクロールして容易に探し出すことができます。デフォルトのセル背景色とフォント ファミリは変更することができます。また、代入されたデータを太字で表示することもできます。

**分析出力。**このグループは、多重代入されたデータセットが分析される際に常に生成されるビューア出力の種類を制御します。デフォルトでは、出力は代入前の元のデータセットと代入された各データセットに対して生成されます。また、代入されたデータのプールをサポートする手順では、最終的なプールされた結果が生成されます。1 変数プールが実行される場合は、プールの診断も表示されます。確認する必要のない出力を無効にすることもできます。

**多重代入オプションを設定するには**

メニューから次の項目を選択します。

[編集] > オプション

[多重代入] タブをクリックします。

# パート II: 例



# 欠損値分析

## 欠損データのパターンの記述

あるデータ通信プロバイダは、顧客データベースにあるサービス利用パターンをさらに良く理解したいと考えています。その企業は、さらに詳細な分析を実行する前に、データが完全に無作為に欠損していることを確認したいと考えています。

顧客データベースから無作為抽出したサンプルは、telco\_missing.sav にあります。詳細は、A 付録 サンプル ファイル in IBM SPSS Missing Values 21 を参照してください。

## 分析を実行して記述統計量を表示

- ▶ 欠損値分析を実行するには、メニューから次の項目を選択します。  
[分析] > [欠損値分析]

図 4-1  
[欠損値分析] ダイアログ ボックス

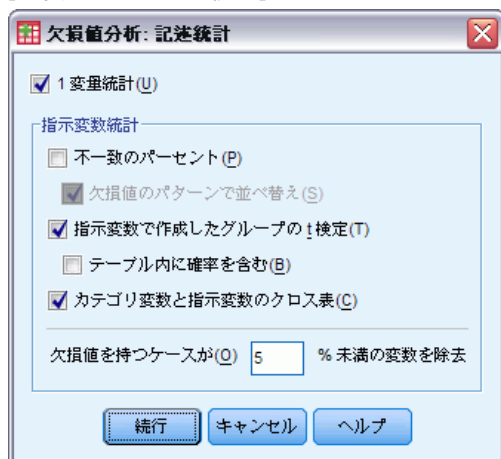


- ▶ カテゴリ変数として、「婚姻状況 [婚姻状況]」、「教育のレベル [教育]」、「退職 [退職]」、および「性別 [性別]」を選択します。
- ▶ 量的（スケール）変数として「サービス月数 [期間]」から「世帯の人数 [世帯人数]」までを選択します。

この時点で手続きを実行して 1 変量統計を得ることはできますが、ここでは追加の記述統計を選択します。

- ▶ [記述統計] をクリックします。

図 4-2  
[欠損値分析: 記述統計] ダイアログ ボックス



[記述統計] ダイアログ ボックスでは、出力に表示するさまざまな記述統計量を指定できます。デフォルトの 1 変量統計量は、欠損データの大まかな範囲を判断するのに役立ちますが、指示変数統計量では、1 つの変数における欠損データのパターンが、どのようにその他の変数値へ影響するかについての詳細がわかります。

- ▶ [指示変数で形成されるグループでの t 検定] を選択します。
- ▶ [カテゴリ変数と指示変数のクロス集計表] を選択します。
- ▶ [続行] をクリックします。
- ▶ [欠損値分析] メイン ダイアログ ボックスの [OK] をクリックします。

## 記述統計の評価

今回の例では、出力に次の項目が含まれます。

- 1 変量統計
- その他の変数が存在または欠損しているときのサブグループの平均値を含む、個別分散 t 検定のテーブル
- それぞれの量的（スケール）変数で各カテゴリの欠損データの度数を表示する、カテゴリ変数ごとのテーブル

図 4-3  
1 変量統計テーブル

	ケースの数	平均値	標準偏差	欠損		極値の数 <sup>a</sup>	
				度数	パーセント	低	高
サービスの継続期間	968	35.56	21.268	32	3.2	0	0
年齢	975	41.75	12.573	25	2.5	0	0
現在の住所	850	11.47	9.965	150	15.0	0	9
収入	821	71.1462	83.14424	179	17.9	0	71
現職の在籍年数	904	11.00	10.113	96	9.6	0	15
世帯人数	966	2.32	1.431	34	3.4	0	33
婚姻状況	885			115	11.5		
教育レベル	965			35	3.5		
退職状況	916			84	8.4		
性別	958			42	4.2		

a. 範囲外のケースの数 (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).

1 変量統計では、変数ごとに欠損データの範囲の外観がわかります。各変数の非欠損値の数は [N] 列に表示され、欠損値の数は [欠損数] 列に表示されます。[欠損パーセント] 列は、欠損値があるケースのパーセンテージを表示し、変数間で欠損データの範囲を比較するのに良い尺度になります。[収入 (家族全体の収入)] で欠損値があるケースが最大 (17.9%) となっており、[年齢 (年齢)] で最小 (2.5%) となっています。[収入] には、極値も最大数あります。

図 4-4  
個別分散 t 検定テーブル

	サレシの推定期間	年齢	現在の住所	収入	現在の在籍年数	世帯人数
現在の住所	.4 202.2	.3 192.5	.	3.5 313.6	1.4 191.1	1.0 199.5
度自由	819	832	850	693	766	824
款効有	149	143	0	128	138	142
款損欠	35.68	41.79	11.47	74.0779	11.20	2.34
有効平均値	34.91	41.49	.	55.2734	9.86	2.21
欠損平均値						
収入	-5.0 249.5	-9.3 222.8	-3.9 191.1	.	-5.9 203.3	3.6 315.2
度自由	793	801	693	821	741	792
款効有	175	174	157	0	163	174
款損欠	33.93	40.01	10.67	71.1462	9.91	2.39
有効平均値	42.97	49.73	14.97	.	15.93	2.02
欠損平均値						
現在の在籍年数	-1.0 110.5	-4 110.2	-7 97.6	.5 114.9	.	-3 110.9
度自由	877	881	766	741	904	874
款効有	91	94	84	80	0	92
款損欠	35.34	41.69	11.37	71.4953	11.00	2.31
有効平均値	37.70	42.27	12.32	67.9125	.	2.37
欠損平均値						
婚姻状況	.0 148.1	1.8 149.5	1.2 138.8	-8 121.2	.9 128.3	-2.2 134.2
度自由	856	862	748	728	805	857
款効有	112	113	102	93	99	109
款損欠	35.56	42.00	11.61	70.3887	11.10	2.28
有効平均値	35.57	39.85	10.43	77.0753	10.17	2.61
欠損平均値						
退職状況	-6 95.4	-4 94.4	-4 84.0	.3 93.2	.	.2 99.0
度自由	888	893	777	751	904	885
款効有	80	82	73	70	0	81
款損欠	35.44	41.70	11.42	71.3356	11.00	2.32
有効平均値	36.89	42.29	11.96	69.1143	.	2.30
欠損平均値						

個別分散 t 検定テーブルは、量的（スケール）変数に影響を与えている 欠損値のパターンが、どの変数であるかを特定するのに役立ちます。t 検定は、個々のケースで変数が存在するか欠損しているかを特定する指示変数を

使用して計算します。指示変数のサブグループの平均値も表にします。最低5%のケースで変数に欠損値がある場合にのみ、指示変数が作成されます。

年齢の高い回答者は、収入レベルを報告しない傾向にあることが示されています。[収入]が欠損している場合の平均の[年齢]は49.73であり、それに対して、[収入]が非欠損である場合は40.01となっています。実際に、[収入]の欠損は、複数の量的（スケール）変数の平均に影響しているように思われます。これは、データが完全に無作為に欠損していないという指標の一つです。

図 4-5  
婚姻状況 [婚姻] のクロス集計表

			合計	未婚	既婚	欠損 % 欠損値
現在の住所	有効	度数	850	390	358	102
		%-セット	85.0	85.5	83.4	88.7
収入	欠損	双欠損値の%	15.0	14.5	16.6	11.3
	有効	度数	821	380	348	93
現職の在籍年数		%-セット	82.1	83.3	81.1	80.9
	欠損	双欠損値の%	17.9	16.7	18.9	19.1
退職状況	有効	度数	904	418	387	99
		%-セット	90.4	91.7	90.2	86.1
	欠損	双欠損値の%	9.6	8.3	9.8	13.9
	有効	度数	916	423	392	101
		%-セット	91.6	92.8	91.4	87.8
	欠損	双欠損値の%	8.4	7.2	8.6	12.2

カテゴリ変数と指示変数のクロス集計表から、個別分散 t 検定テーブルにあるものと類似していることがわかります。指示変数を再度作成すると、このとき以外は、各カテゴリ変数についての全カテゴリにおける度数の計算に使用されます。その値は、欠損値についてカテゴリ間で差があるかどうかの判断に役立ちます。

[婚姻（婚姻状況）]のテーブルを見ると、指示変数内の欠損値数は[婚姻]カテゴリ間であまり変動していないことがわかります。ある人が既婚か未婚であるかは、いずれの量的（スケール）変数についても、データが欠損することに影響がないように思われます。たとえば、未婚の人は、85.5%が[居住年数（現住所での居住年数）]を報告しており、既婚の人は83.4%が報告しています。この差は非常に小さく、偶然によるものと思われます。

図 4-6  
教育レベルのクロス集計表

			合計	高校未満	高校卒	短大卒	4年生・大学卒	大学院卒	欠損 値 の 欠損値
現在の住所	有効	度数	850	163	240	175	186	56	30
		パーセント	85.0	83.2	85.7	88.4	81.9	87.5	85.7
収入	欠損	双対欠損値の%	15.0	16.8	14.3	11.6	18.1	12.5	14.3
		有効	度数	821	155	229	165	193	50
		パーセント	82.1	79.1	81.8	83.3	85.0	78.1	82.9
		双対欠損値の%	17.9	20.9	18.2	16.7	15.0	21.9	17.1
現職の在籍年数	有効	度数	904	178	254	178	204	60	30
		パーセント	90.4	90.8	90.7	89.9	89.9	93.8	85.7
婚姻状況	欠損	双対欠損値の%	9.6	9.2	9.3	10.1	10.1	6.2	14.3
		有効	度数	885	193	278	148	184	52
		パーセント	88.5	98.5	99.3	74.7	81.1	81.2	85.7
		双対欠損値の%	11.5	1.5	.7	25.3	18.9	18.8	14.3
退職状況	有効	度数	916	180	259	180	207	60	30
		パーセント	91.6	91.8	92.5	90.9	91.2	93.8	85.7
	欠損	双対欠損値の%	8.4	8.2	7.5	9.1	8.8	6.2	14.3

次に、[教育（教育レベル）]のクロス集計表について考察します。回答者が少なくともなんらかの大学教育を受けている場合は、婚姻状況の回答について欠損が多い傾向にあります。大学教育を受けていない回答者の少なくとも98.5%は婚姻状況を報告しています。一方、大卒で婚姻状況を報告した人はわずか81.1%でした。これは、何らかの大学教育を受けているが学位がない人よりも低い数字です。

図 4-7  
退職状況 [退職] のクロス集計表

			合計	いいえ	はい	欠損
						有効
現在の住所	有効	度数	850	744	33	73
		パーセント	85.0	85.0	80.5	86.9
収入	欠損	双引欠損値の%	15.0	15.0	19.5	13.1
	有効	度数	821	732	19	70
現職の在籍年数	有効	パーセント	82.1	83.7	46.3	83.3
	欠損	双引欠損値の%	17.9	16.3	53.7	16.7
婚姻状況	有効	度数	904	864	40	0
		パーセント	90.4	98.7	97.6	.0
婚姻状況	欠損	双引欠損値の%	9.6	1.3	2.4	100.0
	有効	度数	885	777	38	70
婚姻状況	有効	パーセント	88.5	88.8	92.7	83.3
	欠損	双引欠損値の%	11.5	11.2	7.3	16.7

[退職 (退職状況)] では、さらに大きな差があることが確認できます。退職者は非退職者に比べて、収入を報告しない傾向がかなりあります。退職している顧客のうち収入レベルを報告したのはわずか 46.3% であり、一方、非退職者で収入レベルを報告している割合は 83.7% でした。



図 4-8  
性別 [性別] のクロス集計表

			合計	男性	女性	欠損
						欠損
現在の住所	有効	度数	850	363	456	31
		割合	85.0	78.6	91.9	73.8
	欠損	汎用欠損値の%	15.0	21.4	8.1	26.2
収入	有効	度数	821	381	406	34
		割合	82.1	82.5	81.9	81.0
	欠損	汎用欠損値の%	17.9	17.5	18.1	19.0
現職の在籍年数	有効	度数	904	412	457	35
		割合	90.4	89.2	92.1	83.3
	欠損	汎用欠損値の%	9.6	10.8	7.9	16.7
婚姻状況	有効	度数	885	400	445	40
		割合	88.5	86.6	89.7	95.2
	欠損	汎用欠損値の%	11.5	13.4	10.3	4.8
退職状況	有効	度数	916	420	461	35
		割合	91.6	90.9	92.9	83.3
	欠損	汎用欠損値の%	8.4	9.1	7.1	16.7

[性別 (性別)] では別の相違点があります。住所情報は女性よりも男性に欠損が多くみられます。この相違は偶然である可能性もありますが、そうではないように思われます。データは完全に無作為に欠損しているようには見えません。

欠損データのパターンに注目して詳細を検討します。

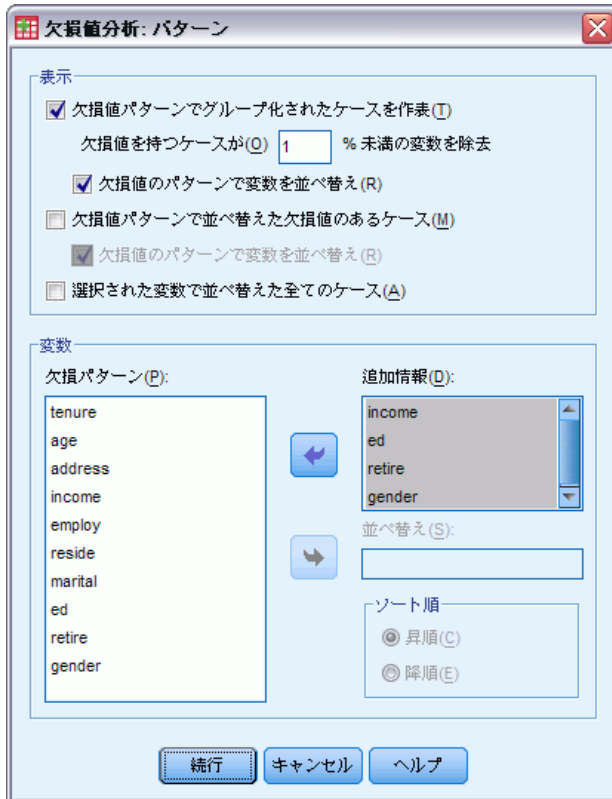
## パターンを表示するために分析を再実行

図 4-9  
[欠損値分析] ダイアログ ボックス



- ▶ [欠損値分析のリコール] ダイアログ ボックス。このダイアログは、前回の分析で使用した変数を記憶しています。それを変更しないでください。
- ▶ [パターン] をクリックします。

図 4-10  
[Missing Value Analysis Patterns (欠損値分析 - パターン)] ダイアログ ボックス



[パターン] ダイアログ ボックスでは、さまざまなパターンのテーブルを選択できます。欠損値パターンでグループ化され、集計されたパターンを表示します。[教育 (教育レベル)]、[退職 (退職状況)]、および [性別 (性別)] での欠損パターンは、データに影響を与えたように思われるので、これらの変数の追加情報の表示を選択します。欠損値が非常に多いという理由から、[収入 (家族全体の収入)] の追加情報も含めます。

- ▶ [欠損値パターンでグループ化され、集計されたケース] を選択します。
- ▶ [収入]、[教育]、[退職]、および [性別] を選択し、[追加情報] リストに追加します。
- ▶ [続行] をクリックします。
- ▶ [欠損値分析] メイン ダイアログ ボックスの [OK] をクリックします。

## パターン テーブルの評価

図 4-11  
集計されたパターン テーブル

ケースの 数	欠損パターン <sup>a</sup>									完了 数	収入	教育レベル <sup>d</sup>					退職状況 <sup>d</sup>		性別 <sup>d</sup>		
	年齢	世帯人数	サ ー ビ ス の 継 続 期 間	教育 レベル	性別	退職 状況	現職の 在籍年数	婚姻 状況	現在の 住所			収入	高校未 満	高校 卒	短大 卒	専 修 学 校 卒	大 学 院 卒	い い え	は い	男 性	女 性
475										X	475	76.5853	99	157	87	101	31	463	12	201	274
109										X	584	.	27	35	19	17	11	95	14	47	62
16								X		X	687	.	5	9	0	1	1	12	4	12	4
87		X						X			562	54.4368	21	27	9	24	6	85	2	66	21
13			X								488	56.0000	4	3	2	3	1	13	0	4	9
60							X				535	77.2167	1	2	27	24	6	59	1	35	25
16				X							491	47.8125	0	0	0	0	0	16	0	6	10
17			X								492	76.2353	2	7	3	4	1	17	0	7	10
18					X						493	54.1111	3	7	4	4	0	17	1	0	0
16							X		X		660	.	0	0	7	8	1	14	2	6	10
37					X	X					520	59.4595	9	14	5	8	1	0	0	15	22

1% 未満のケースをもつパターン (10 以下) は表示されません。

a. 変数は欠損パターンをもとに並べ替えられています。

b. このパターンで欠損している変数 (X とマークされている) が使用されない場合に完了するケースの数。

c. 各々の独自のパターンにおける平均値

d. 各々の独自のパターンにおける度数分布

集計されたパターン テーブルでは、個々のケースについて、複数の変数でデータ欠損が出る傾向にあるかを表示します。これは、データが同時に欠損しているかどうかの判断に役立ちます。

1% 以上のケースで同時に欠損データが発生しているパターンは 3 つあります。変数 [勤続年数 (現職の雇用期間)] および [退職 (退職状況)] は共に、他のペアより欠損している場合がよくあります。[退職状況] および [勤続年数] の記録は類似している情報なので、これは驚くことではありません。回答者が退職しているかどうかわからない場合は、回答者の勤続年数もおそらくわからないでしょう。

[収入 (家族全体の収入)] の平均は、欠損値パターンによってかなり変動しているように思われます。特に、[婚姻 (婚姻状況)] が欠損している場合に、[収入] の平均は、このケースで 6% (1000 件中 60 件) も高くなっています ([期間 (サービス月数)] が欠損している場合も高くなっていますが、このパターンはわずか 1.7% のケースにしかりません)。より高いレベルの教育を受けた人は、婚姻状況についての質問に回答しない傾向があったことを思い出してください。[教育 (教育レベル)] に示された度数で、この傾向がわかります。より高いレベルの教育を受けた人は、より高い収入を得て婚姻状況を報告しない傾向にあると仮定することで、[収入] の増加について説明できるかもしれません。

記述統計と欠損データのパターンを考慮すると、データは完全に無作為に欠損していないと結論付けることができます。EM 推定値と併せて表示されている Little の MCAR 検定を通じて、この結論を確認できます。

## Little の MCAR 検定のために分析を再実行

図 4-12  
[欠損値分析] ダイアログ ボックス



- ▶ [欠損値分析のリコール] ダイアログ ボックス。
- ▶ [EM] をクリックします。
- ▶ [OK] をクリックします。

図 4-13  
EM 平均値テーブル

サービスの継続期間	年齢	現在の住所	収入	現職の在籍年数	世帯人数
36.12	41.91	11.58	77.3941	11.22	2.29

a. Little の MCAR 検定: カイ 2 乗 = 179.836, DF = 107, Sig. = .000

Little の MCAR 検定の結果は、それぞれの EM 推定値テーブルの脚注に表示されます。Little の MCAR 検定の帰無仮説は、データが完全に無作為に欠損している (MCAR) ことです。欠損値パターンがデータ値によらない場合は、データは MCAR です。このサンプルでは有意確率が 0.05 未満なので、

データは完全に無作為に欠損していないと結論付けることができます。これによって、記述統計と集計したパターンから得た結論を確認できます。

この点で、データは完全に無作為に欠損していないため、リストごとに欠損値を持つケースを削除、または欠損値を単独で代入することは適切ではありません。ただし、[複数代入](#)を使用して、データセットをより詳細に分析できます。

# [多重代入]

## 多重代入を使用したデータセットの比較と分析

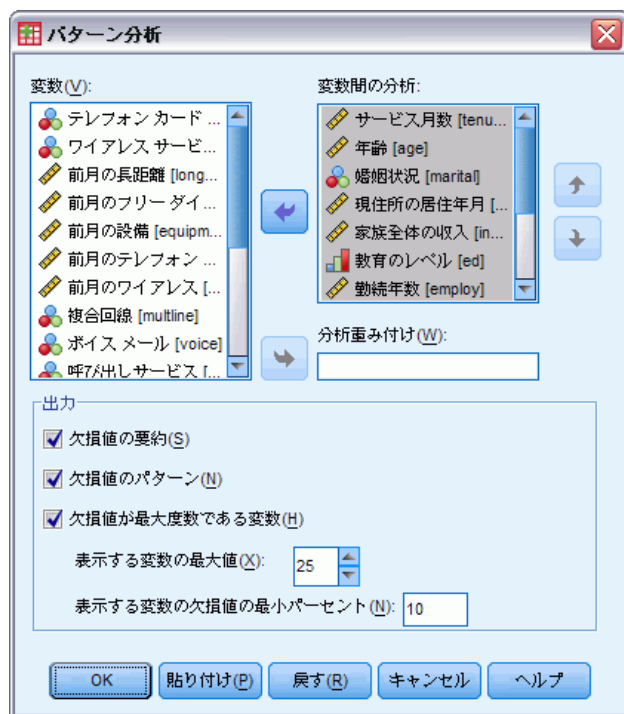
あるデータ通信プロバイダは、顧客データベースにあるサービス利用パターンをさらに良く理解したいと考えています。彼らは、顧客が使用するサービスの完全なデータセットを持っていますが、この企業が収集した人口統計情報には、多くの欠損値が存在します。さらに、この値は完全にランダムで欠損しているわけではないため、多重代入を利用して、データセットを完成させます。

顧客データベースから無作為抽出したサンプルは、telco\_missing.sav にあります。詳細は、[A 付録 サンプル ファイル in IBM SPSS Missing Values 21](#) を参照してください。

## 欠損値のパターンの分析

- ▶ 第一段階として、欠損値のパターンを調べます。メニューから次の項目を選択します。  
[分析] > [多重代入] > [パターン分析]

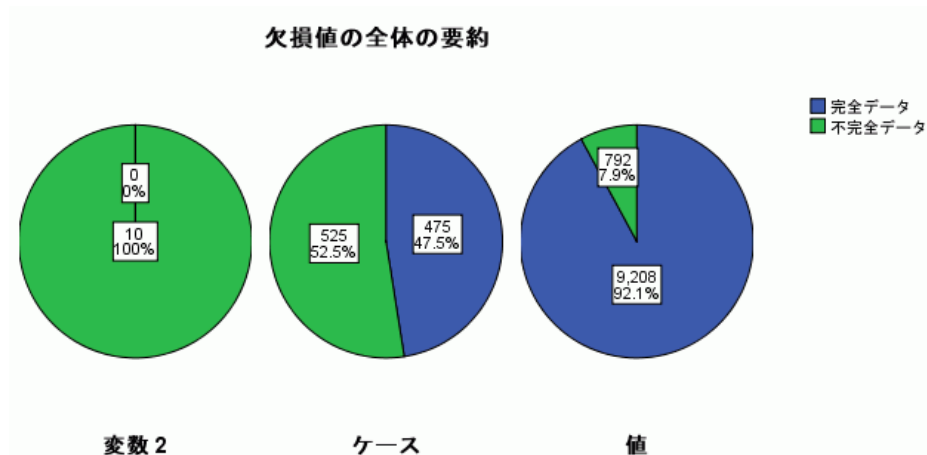
図 5-1  
[パターン分析] ダイアログ



- ▶ 分析変数として「サービス月数 [期間]」から「世帯の人数 [世帯人数]」までを選択します。

## 全体の要約

図 5-2  
欠損値の全体の要約





欠損値の全体の要約は、データの欠損値の異なる要素を表す 3 つの円グラフを表示します。

- 変数グラフは、10 個の分析変数ごとに、ケースに少なくとも 1 つの欠損値が含まれることを示しています。
- ケース グラフは、1000 ケース中の 525 ケースが、変数に少なくとも 1 つの欠損値を持つことを示しています。
- 値グラフは、1000 の値中の 792 個（ケース × 変数）が欠損していることを示しています。

欠損値を含むケースは、平均で、変数 10 個につき、約 1.5 個の欠損値を持っています。これは、リストごとに削除を行うと、データセットの情報の多くが失われることを意味しています。

## 変数の要約

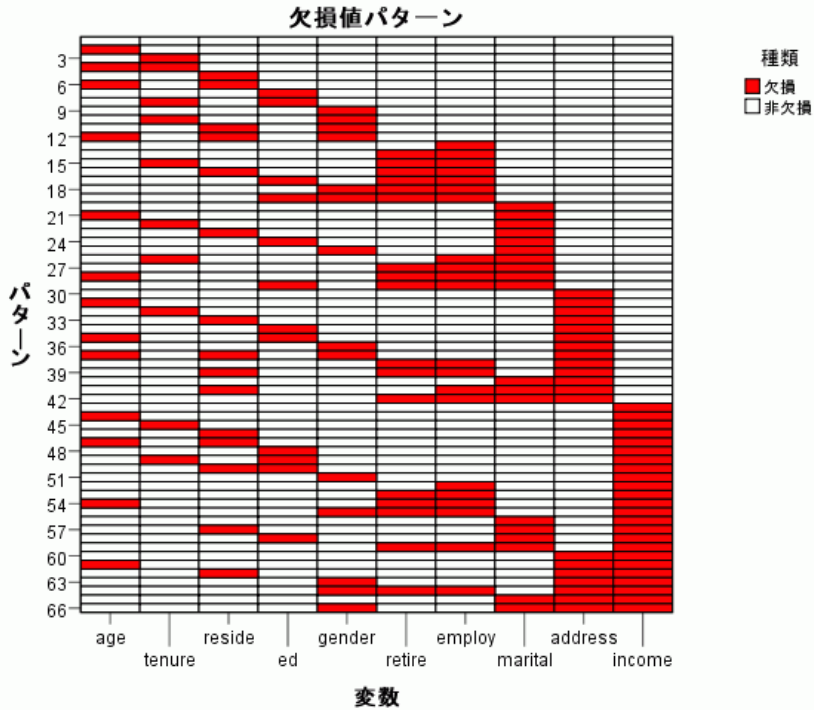
図 5-3  
変数の要約

	欠損		有効な N	平均値 (ラン検定)	標準偏差
	度数	パーセント			
家族全体の収入	179	17.9%	821	71.1462	83.14424
現住所の居住年月	150	15.0%	850	11.47	9.965
婚姻状況	115	11.5%	885		

変数の要約は、少なくとも 10% の欠損値を含む変数として表示され、テーブルのそれぞれの変数の欠損値の数とパーセントを示します。さらに、スケール変数の有効な値の平均値と標準偏差と、すべての変数の有効な値の数を表示します。「世帯全体の収入」、「現住所の居住年月」、および「婚姻状況」は、この順に最も多く欠損値が含まれています。

## パターン

図 5-4  
欠損値パターン



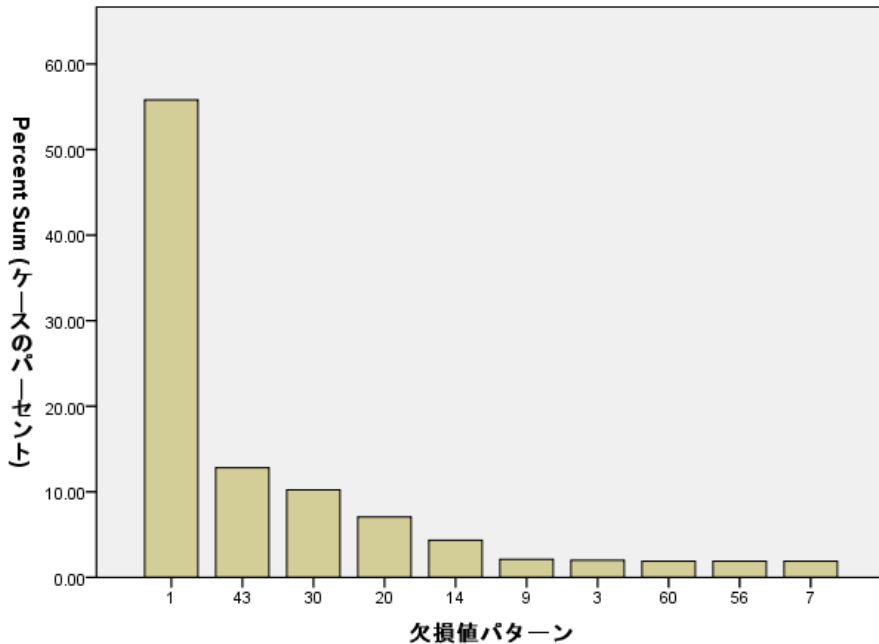
パターン グラフは、分析変数の欠損値パターンを表示します。それぞれのパターンは、不完全および完全なデータの同じパターンを持つケースのグループに対応します。たとえば、パターン 1 は、欠損値のないケースを表し、パターン 33 は、欠損値が「家族人数（家族の人数）」および「居住年数（現住所での居住年数）」に存在するケースを表し、パターン 66 は、欠損値が「性別（性別）」、「婚姻状況（婚姻状況）」、「居住年数」、および「収入（家族全体の収入）」に存在するケースを表します。データセットは潜在的に 2 種類の変数の数パターンになります。10 個の分析変数は  $2^{10}=1024$  ですが、66 パターンのみが、データセットの 1000 ケース内に示されます。

グラフは、単調性を明らかにするために、分析変数とパターンを整理します。具体的には、変数は左から右に欠損値の昇順に並べられます。パターンは、最後の変数によってソートされ（最初に非欠損値、次に欠損値）、次に最後から 2 番目の変数によってソートされるというように、右から左に処理されます。これにより、データに対して単調な代入を使用できるか判定することができます。使用できない場合は、データが単調なパターンにどのくらい近似しているのかが明らかになります。データが単調である場合、グラフのすべての欠損セルと非欠損セルは連続しま

す。つまり、グラフの右下部に非欠損セルの「島」はできず、グラフの左上部に欠損セルの「島」ができたりしません。

このデータセットは単調でなく、単調性を実現するために、代入する必要のある値が多く存在します。

図 5-5  
パターン度数



パターンが要求された場合、対となる棒グラフは、各パターンのケースのパーセントを表示します。これは、データセットのケースの半数以上がパターン 1 であることを示し、欠損値パターン グラムは、これが欠損値を含まないケースのパターンであることを示します。パターン 43 は、「収入」に欠損値を含むケースを表し、パターン 30 は、「居住年数」に欠損値を含むケースを表し、パターン 20 は、「婚姻状況」に欠損値を含むケースを表します。ケースの大部分、およそ 5 分の 4 は、これらの 4 つのパターンで表されます。10 種類の最も頻繁に現われるパターンの中で、複数の変数に欠損値を含むケースを表すのは、パターン 14、60、56 だけです。

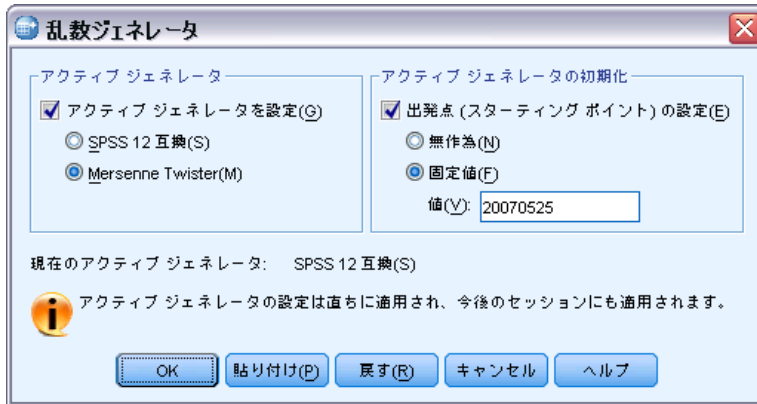
単調方法の使用が実現できない場合を除いて、欠損パターンの分析で、多重代入を行うための特定の障害が明らかになることはありません。

## 欠損値の自動代入

値を代入する用意ができたので、自動設定で実行してみましよう。ただし、代入を要求する前に、乱数シードを設定します。乱数シードを設定すると、分析を正確に複製できます。

- ▶ 乱数シードを設定するには、メニューから次の項目を選択します。  
[変換] > [乱数ジェネレータ...]

図 5-6  
[乱数ジェネレータ] ダイアログ ボックス



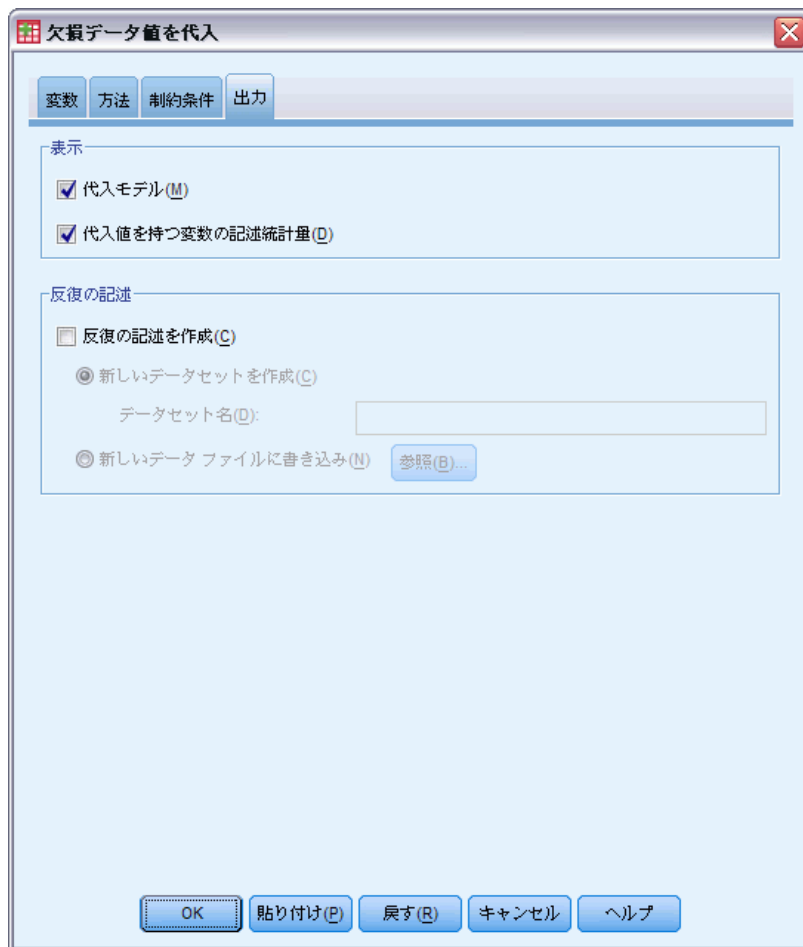
- ▶ [アクティブ ジェネレータを設定] を選択します。
- ▶ [Mersenne Twister] を選択します。
- ▶ [出発点(スターティングポイント)の設定] を選択します。
- ▶ [固定値] を選択し、値として「20070525」と入力します。
- ▶ [OK] をクリックします。
- ▶ 欠損データ値を多重代入するには、メニューから次の項目を選択します。  
[分析] > [多重代入] > [欠損データ値を代入]

図 5-7  
[欠損データ値の代入] ダイアログ



- ▶ 代入モデルの変数として「サービス月数 [期間]」から「世帯の人数 [世帯人数]」までを選択します。
- ▶ 代入データを保存するデータセットとして、「telcoImputed」と入力します。
- ▶ [出力] タブをクリックします。

図 5-8  
[出力] タブ



- ▶ [代入値を持つ変数の記述統計量] を選択します。
- ▶ [OK] をクリックします。

## 代入指定

図 5-9  
代入指定

代入法	自動	
代入数		5
スケール変数のモデル モデルに含まれる交互…	線型回帰 (なし)	
欠損値の最大パーセント		100.0%
Maximum Number of Parameters in Imputati…		100

代入指定テーブルは、指定が正しいにことを確認するために、これまでに  
行った指定を確認するのに役立ちます。

## 代入結果

図 5-10  
代入結果

代入法	完全条件指定	10
完全条件指定法の反復		
従属変数 代入済み	tenure, age, marital, address, income, ed, employ, retire, gender, reside	
未代入 (欠損値が多すぎる)		
未代入 (欠損値がない)		
代入順序	age, tenure, reside, ed, gender, retire, employ, marital, address, income	

代入結果は、代入プロセスにおいて、実際に何が起きているかの概要を  
確認できます。特に次の点に注意する必要があります。

- 指定テーブルの代入方法は [自動] で、自動で実際に選択される方法  
は、[完全条件指定] です。
- 要求された変数はすべて代入されます。
- 代入順序は、変数が欠損値パターン グラフの x-軸に出現する順番  
です。

## 代入モデル

図 5-11  
代入モデル

	モデル		欠損値	代入値
	種類	効果		
年齢	線型回帰	ed,gender,retire,marital,tenure,reside,employ,address,income	25	125
サービス月数	線型回帰	ed,gender,retire,marital,age,reside,employ,address,income	32	160
世帯の人数	線型回帰	ed,gender,retire,marital,age,tenure,employ,address,income	34	170
教育のレベル	ロジスティック回帰	gender,retire,marital,age,tenure,reside,employ,address,income	35	175
性別	ロジスティック回帰	ed,retire,marital,age,tenure,reside,employ,address,income	42	210
退職	ロジスティック回帰	ed,gender,marital,age,tenure,reside,employ,address,income	84	420
勤続年数	線型回帰	ed,gender,retire,marital,age,tenure,reside,address,income	96	480
婚姻状況	ロジスティック回帰	ed,gender,retire,age,tenure,reside,employ,address,income	115	575
現住所の居住年月	線型回帰	ed,gender,retire,marital,age,tenure,reside,employ,income	150	750
家族全体の収入	線型回帰	ed,gender,retire,marital,age,tenure,reside,employ,address	179	895

代入モデル テーブルでは、それぞれの変数がどのように代入されたかを、さらに詳細に確認することができます。特に次の点に注意する必要があります。

- 変数は、代入順序の順でリストアップされます。
- スケール変数は、線型回帰モデルにモデル化され、カテゴリ変数は、ロジスティック回帰モデルにモデル化されます。
- それぞれのモデルは、その他すべての変数を、主効果として使用します。
- 各変数の欠損値の数が、その変数に代入された値の総数とともに報告されます（欠損数 × 代入数）。



## [記述統計]

図 5-12  
期間（サービス月間）の記述統計量

データ	代入	度数	平均値(ラン 検定)	標準偏差	最小値	最大値
元のデータ		968	35.56	21.268	1.00	72.00
代入値	1	32	40.29	26.148	-18.23	88.16
	2	32	41.81	24.211	-9.49	108.92
	3	32	38.42	23.790	-3.02	85.21
	4	32	39.22	24.907	-5.93	88.35
	5	32	38.17	19.410	-2.18	95.55
代入後の完全データ	1	1000	35.71	21.442	-18.23	88.16
	2	1000	35.76	21.383	-9.49	108.92
	3	1000	35.65	21.346	-3.02	85.21
	4	1000	35.68	21.389	-5.93	88.35
	5	1000	35.64	21.207	-2.18	95.55

記述統計量テーブルは、代入値を持つ変数の要約を示します。個々のテーブルは、変数ごとに作成されます。表示される統計量の種類は、スケール変数かカテゴリ変数かによって変わります。

スケール変数の統計量には、元のデータの度数、平均値、標準偏差、最小値、最大値、それぞれの代入値、およびそれぞれの完全なデータセット（元の値と代入値の組み合わせ）が含まれます。

「期間（サービス月間）」の記述統計量は、元のデータの値とほぼ等しい代入値のセットにおける平均値と標準偏差を表します。ただし、最小値で、「期間」にマイナスの値が代入されていることで、明らかな問題があることが分かります。

図 5-13  
婚姻状況の記述統計量

データ	代入	カ...	度数	パーセント
元のデータ		0	456	51.5
		1	429	48.5
代入値	1	0	50	43.5
		1	65	56.5
	2	0	46	40.0
		1	69	60.0
	3	0	55	47.8
		1	60	52.2
	4	0	51	44.3
		1	64	55.7
	5	0	57	49.6
		1	58	50.4
代入後の完全データ	1	0	506	50.6
		1	494	49.4
	2	0	502	50.2
		1	498	49.8
	3	0	511	51.1
		1	489	48.9
	4	0	507	50.7
		1	493	49.3
	5	0	513	51.3
		1	487	48.7

カテゴリ変数では、統計量には、元のデータのカテゴリの度数とパーセント、代入値、および完全なデータが含まれます。「婚姻状況」のテーブルには、面白い結果が含まれています。代入値として、ケースの大部分が、もとのデータより多く結婚しているものと推定されています。これは、ランダムな変動、つまり、欠損の機会がこの変数の値と関連していることに起因する場合があります。

図 5-14  
収入（家族全体の収入）の記述統計量

データ	代入	度数	平均値 (ラン 検定)	標準偏差	最小値	最大値
元のデータ		821	71.1462	83.14424	9.0000	944.0000
代入値	1	179	90.7360	101.27724	-140.8198	333.2438
	2	179	93.6489	93.87048	-123.7850	370.7097
	3	179	101.3281	103.85259	-240.3800	408.9143
	4	179	84.7149	93.14576	-199.2433	335.0470
	5	179	98.0255	91.20978	-92.3070	355.5226
代入後の完全データ	1	1000	74.6528	86.93875	-140.8198	944.0000
	2	1000	75.1741	85.55023	-123.7850	944.0000
	3	1000	76.5487	87.92057	-240.3800	944.0000
	4	1000	73.5750	85.13099	-199.2433	944.0000
	5	1000	75.9576	85.22261	-92.3070	944.0000

「期間」と同様に、その他すべてのスケール変数、「収入（家族全体の収入）」はマイナスの代入値 - を示しており、特定の変数に制約条件を設けてユーザー指定によるモデルを実行する必要があります。しかし、「収入」に、別の潜在的な問題があります。代入ごとの平均値が元のデータよりかなり高く、代入ごとの最大値が元のデータよりかなり低くなっています。収入の分布は、かなり右に歪んだ傾向にあり、したがって、これが問題の原因である可能性があります。

## ユーザー指定による代入モデル

代入値が各変数の値の妥当な範囲外に出てしまうことを避けるため、変数に制約条件を設けてユーザー指定による代入モデルを指定します。さらに、「家族全体の収入」がかなり右に歪んでおり、詳細な分析に「収入」の対数を使用するのが望ましいため、収入の対数を直接入力しても問題ないと思われれます。

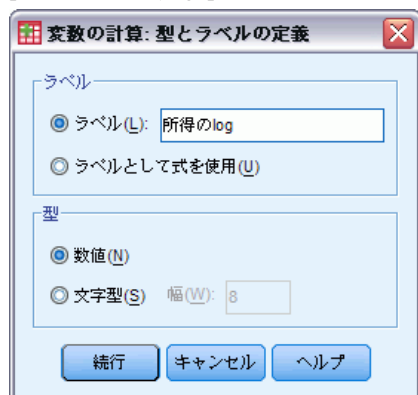
- ▶ 元のデータセットがアクティブであることを確認します。
- ▶ 収入の対数の変数を作成するには、メニューから次の項目を選択します。  
[変換] > [変数の計算...]

図 5-15  
[変数の計算] ダイアログ



- ▶ 目標変数として 「lninc」 と入力します。
- ▶ 数式として「ln(income)」と入力します。
- ▶ [型 & ラベル] をクリックします。

図 5-16  
[型とラベルの定義] ダイアログ ボックス



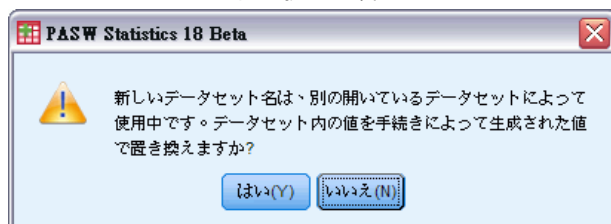
- ▶ ラベルに「Log of income」と入力します。
- ▶ [続行] をクリックします。
- ▶ [変数の計算] ダイアログ ボックスで [OK] をクリックします。

図 5-17  
代入モデルの家族全体の収入を置き換える「Log of income」という[変数]タブ



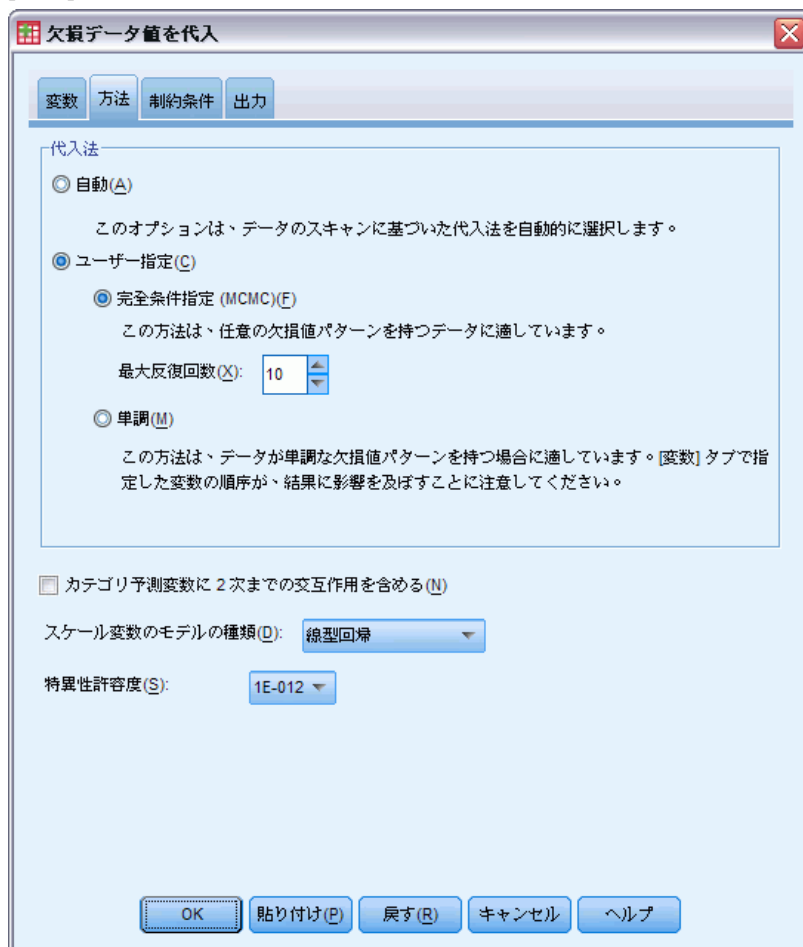
- ▶ [欠損データ値を代入] ダイアログ ボックスをもう一度開き、[変数] タブをクリックします。
- ▶ [家族全体の収入 [収入]] のチェックを外し、モデルの変数として [Log of income [lninc]] を選択します。
- ▶ [方法] タブをクリックします。

図 5-18  
既存のデータセットの置き換えの警告



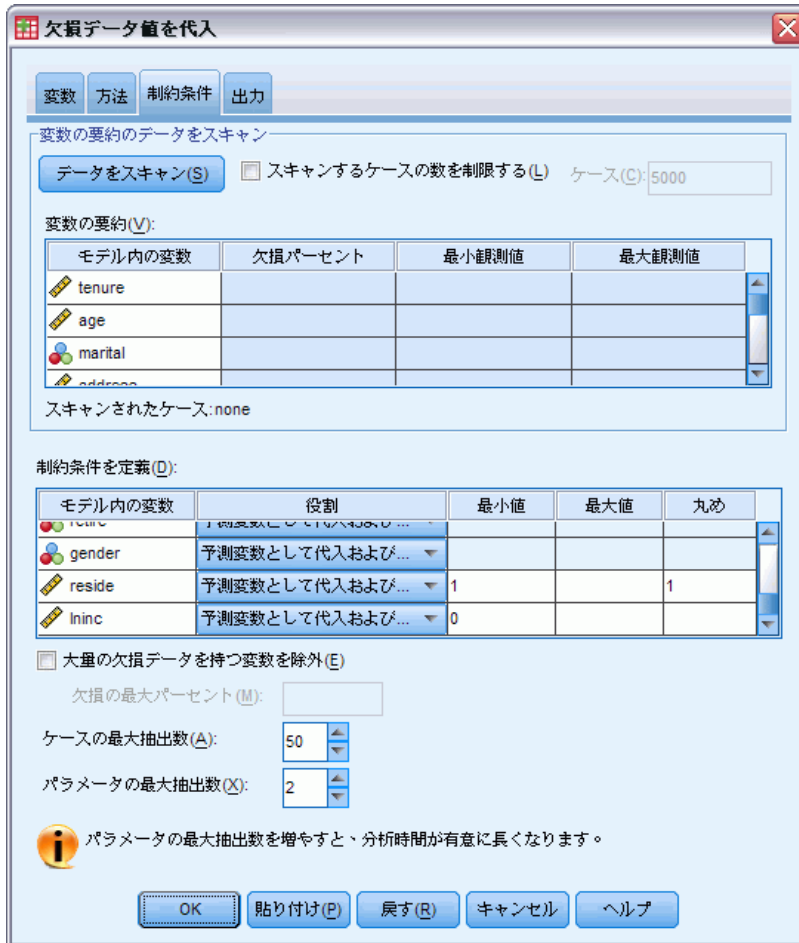
- ▶ 警告が表示された場合は、[はい] をクリックします。

図 5-19  
[方法] タブ



- ▶ [ユーザー指定] を選択し、代入方法として、[完全条件指定] を選択したままの状態にします。
- ▶ [制約条件] タブをクリックします。

図 5-20  
[制約条件] タブ

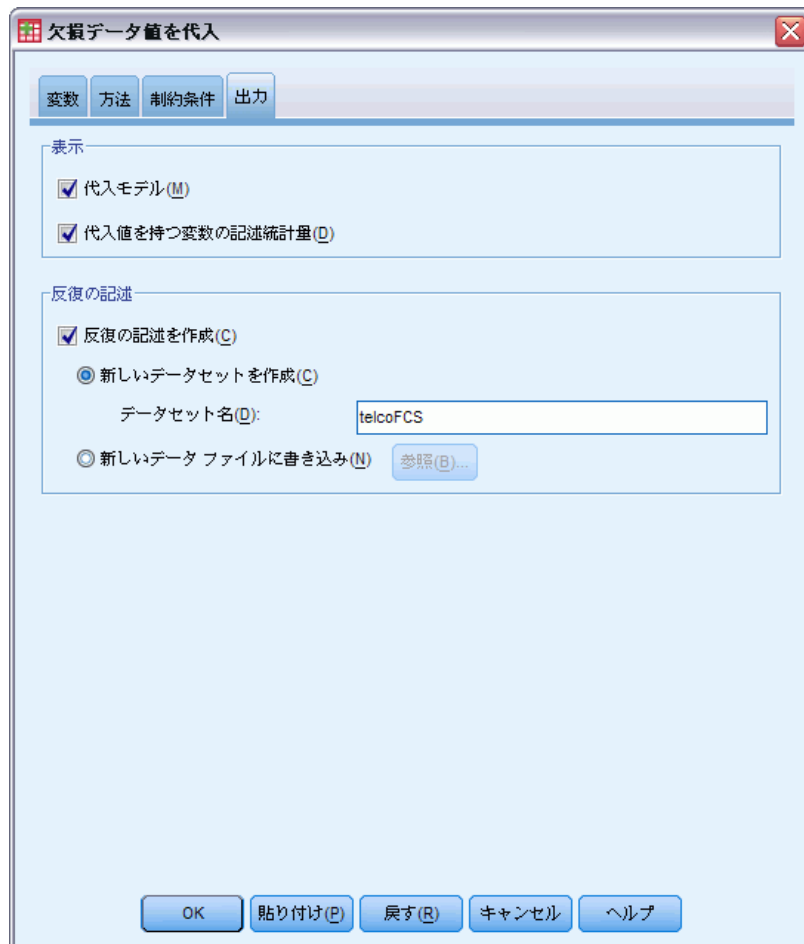


- ▶ [データのスキャン] をクリックします。
- ▶ [制約の定義] グリッドで、「サービス年間[期間]」の最小値に「1」と入力します。
- ▶ 「年齢」の最小値として「18」と入力します。.
- ▶ 「居住年数(現住所での居住年数)」の最小値として「0」と入力します。.
- ▶ 「雇用(勤続年数)」の最小値として「0」と入力します。
- ▶ 「同居人数(世帯の同居人数)」の最小値として「1」、丸めとして「1」と入力します。他のスケール変数の多くは、整数値として報告されますが、現在の住所に誰かが 13.8 年間住んでいると仮定することはできますが、2.2 人がそこに住んでいるとは実際は考えられません。



- ▶ 「lninc (収入の対数)」の最小値として「0」と入力します。
- ▶ [出力] タブをクリックします。

図 5-21  
[出力] タブ



- ▶ [反復の記述を作成] を選択し、新しいデータセット名として、「telcoFCS」と入力します。
- ▶ [OK] をクリックします。

## 代入制限

図 5-22  
代入制限

	代入での役割		代入値		
	従属変数	予測	最小値	最大値	丸め
サービス月数	はい	はい	1	(なし)	
年齢	はい	はい	18	(なし)	
婚姻状況	はい	はい			
現住所の居住年月	はい	はい	0	(なし)	
教育のレベル	はい	はい			
勤続年数	はい	はい	0	(なし)	
退職	はい	はい			
性別	はい	はい			
世帯の人数	はい	はい	1	(なし)	整数
ログの所得	はい	はい	0	(なし)	

ユーザー指定の代入モデルは、結果として、代入モデルに設けられた制約を確認する新しいテーブルになります。すべては、ユーザーの指定にしたがって、作成されます。

## [記述統計]

図 5-23  
期間 (サービス月間) の記述統計量

データ	代入	度数	平均値 (ラン検定)	標準偏差	最小値	最大値
元のデータ		968	35.56	21.268	1.00	72.00
代入値	1	32	38.93	18.327	9.66	65.94
	2	32	36.78	22.577	2.73	86.82
	3	32	41.17	24.372	7.63	102.97
	4	32	37.15	20.717	2.15	86.30
	5	32	40.04	25.375	2.20	93.41
代入後の完全データ	1	1000	35.67	21.181	1.00	72.00
	2	1000	35.60	21.300	1.00	86.82
	3	1000	35.74	21.383	1.00	102.97
	4	1000	35.61	21.242	1.00	86.30
	5	1000	35.70	21.411	1.00	93.41

制約条件が設けられたユーザー指定の代入モデルの「期間 (サービス月間)」の記述統計量テーブルは、「期間」に、マイナスの値が代入される問題が解決していることが分かります。

図 5-24  
婚姻状況の記述統計量

データ	代入	カ...	度数	パーセント
元のデータ		0	456	51.5
		1	429	48.5
代入値	1	0	39	33.9
		1	76	66.1
	2	0	49	42.6
		1	66	57.4
	3	0	57	49.6
		1	58	50.4
	4	0	55	47.8
		1	60	52.2
	5	0	52	45.2
		1	63	54.8
代入後の完全データ	1	0	495	49.5
		1	505	50.5
	2	0	505	50.5
		1	495	49.5
	3	0	513	51.3
		1	487	48.7
	4	0	511	51.1
		1	489	48.9
	5	0	508	50.8
		1	492	49.2

「婚姻状況」のテーブルは、分布が元のデータにさらに一致した代入 (3) が行われていますが、大部分は依然として、元のデータよりも多くのケースが婚姻状態にあることを示しています。これは、ランダムな変動が原因であることが考えられますが、これらの値がランダムに欠損 (MAR) していないかどうかを決定するために、詳細にデータを研究する必要があるかもしれません。ここでは、これ以上追求することはしません。

図 5-25  
lninc (収入の対数) の記述統計量

データ	代入	度数	平均値 (ラン 検定)	標準偏差	最小値	最大値
元のデータ		821	3.9291	.75305	2.1972	6.8501
代入値	1	179	4.2708	.89640	2.0729	6.6506
	2	179	4.1911	.86732	2.1596	6.6504
	3	179	4.1387	.88927	2.0088	6.2921
	4	179	4.2412	.91367	2.1202	6.6517
	5	179	4.1635	.93344	2.1523	6.7891
代入後の完全データ	1	1000	3.9902	.79109	2.0729	6.8501
	2	1000	3.9760	.78077	2.1596	6.8501
	3	1000	3.9666	.78284	2.0088	6.8501
	4	1000	3.9849	.79281	2.1202	6.8501
	5	1000	3.9710	.79297	2.1523	6.8501

「期間」と同様、その他すべてのスケール変数、「lninc (収入の対数)」は、マイナスの代入値を表しません。さらに、代入の平均値は、-「収入」スケールの自動代入の実行時より元のデータの平均値に近く、「lninc」の元のデータの平均値は、約  $e^{3.9291}=50.86$  です。それに対し、代入の標準的な平均値は、約  $e^{4.2}=66.69$  です。また、それぞれの代入の最大値は、元のデータの最大値に近くなっています。

## FCS 収束の確認

完全条件指定代入方法を使用する場合、反復による平均値と標準偏差のプロットおよび、モデル収束の評価に利用するために代入される値のスケール従属変数ごとの代入を確認するのが良いです。

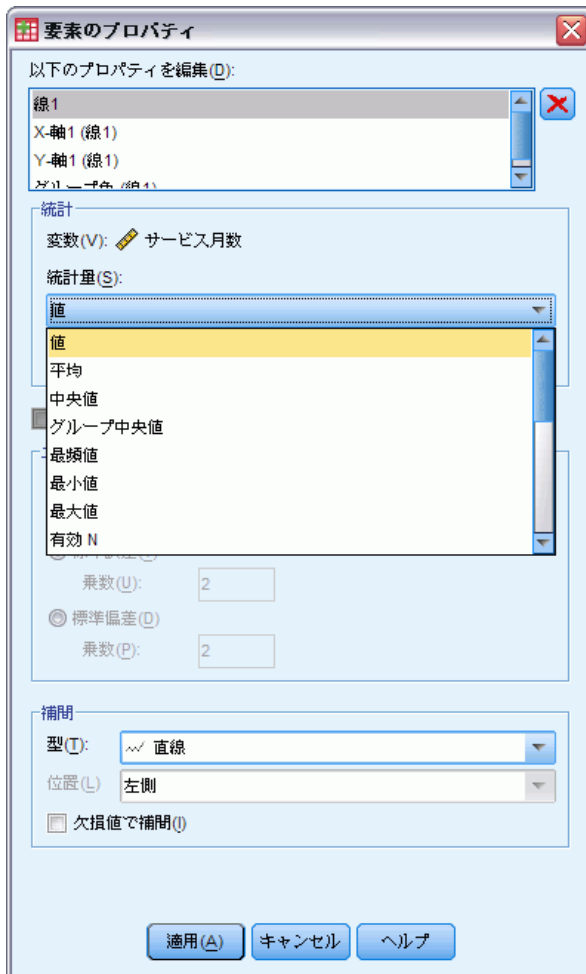
- ▶ この種類の図表を作成するには、telcoFCS データセットをアクティブにし、メニューから次の項目を選択します。  
[グラフ] > [図表ビルダー...]

図 5-26  
図表ビルダー、複数の線のプロット



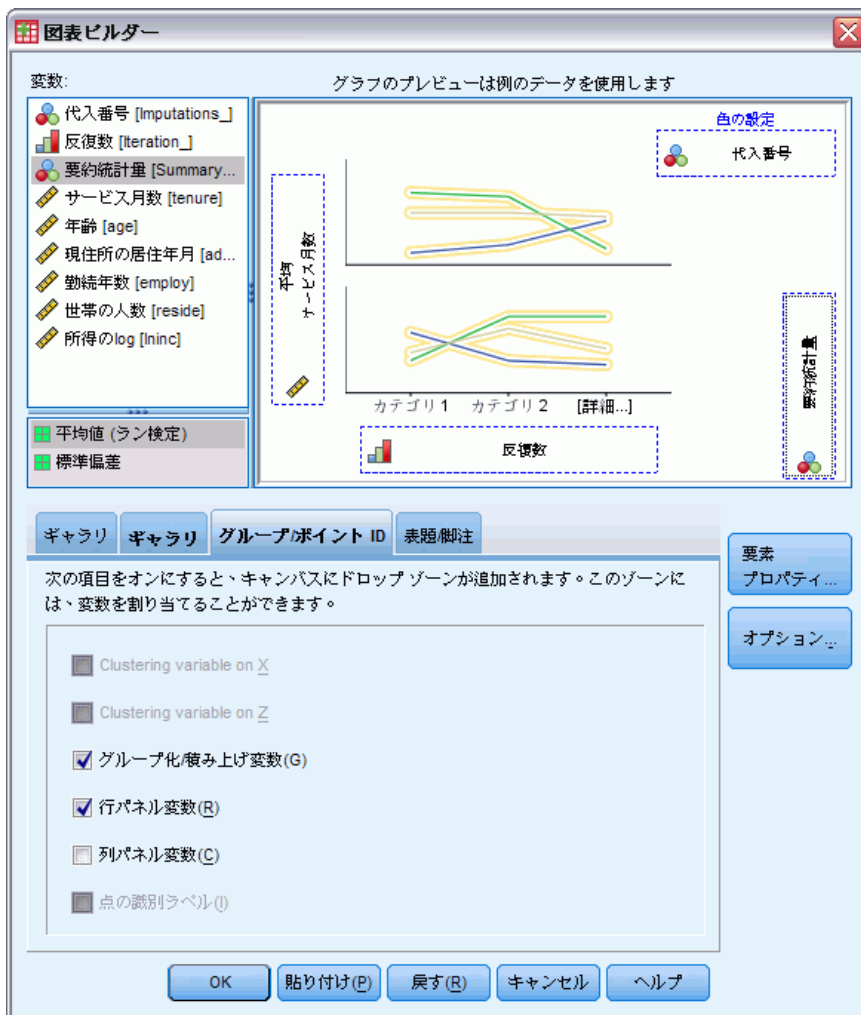
- ▶ [線] ギャラリーを選択し、[複数の線] を選択します。
- ▶ 「サービス月間 [期間]」を、Y 軸にプロットする変数として選択します。
- ▶ X 軸にプロットする変数として、[反復回数[Iteration\_]] を選択します。
- ▶ 色を設定する変数として [反復回数[Iteration\_]] を選択します。

図 5-27  
図表ビルダー、要素のプロパティ



- ▶ 要素のプロパティで、表示する統計量として [値] を選択します。
- ▶ [適用] をクリックします。
- ▶ 図表ビルダーで、[グループ/ポイント ID] タブをクリックします。

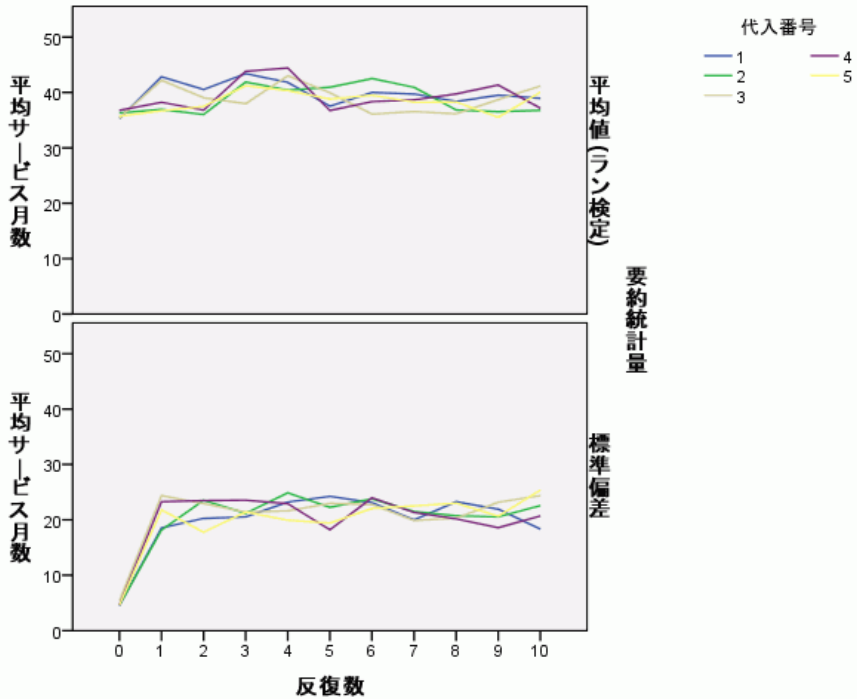
図 5-28  
図表ビルダー、[グループ/ポイント ID] タブ



- ▶ [行パネル変数] を選択します。
- ▶ パネル変数として、[要約統計量 [SummaryStatistic\_]] を選択します。
- ▶ [OK] をクリックします。

## FCS 収束グラフ

図 5-29  
FCS 収束グラフ



複数の線グラフを 1 組作成して、5 回の要求された代入それぞれに、FCS 代入方法の反復ごとの「サービス月間 [期間]」の代入値の平均値と標準偏差を表示しました。このプロットの目的は、線のパターンを確認することです。パターンは無く、「ランダム」であるようです。他のスケール変数についても、同様のプロットを作成することができますが、そのプロットも、認識できるパターンを示しません。

## 完成したデータの分析

代入値が、満足のいくものであるので、「完成した」データの分析を実行してみましょう。データセットには、顧客ベースをサービスの使用パターンによって区分する変数「顧客カテゴリ [custcat]」があり、顧客を 4 つのグループに分類します。顧客がどのグループに属するかを、人口統計情報を使用したモデルを適用して予測できれば、個々の見込み客にあわせてサービスをカスタマイズすることができます。



- ▶ telcoImputed データセットをアクティブにします。完成したデータの多項ロジスティック回帰モデルを作成するには、メニューから次の項目を選択します。

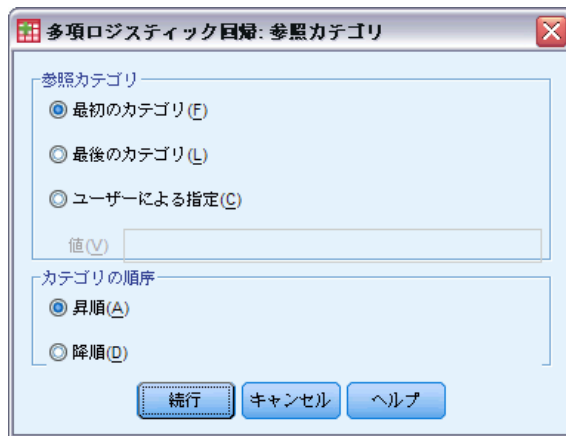
[分析] > [回帰] > [多項ロジスティック...]

図 5-30  
[多項ロジスティック回帰] ダイアログ



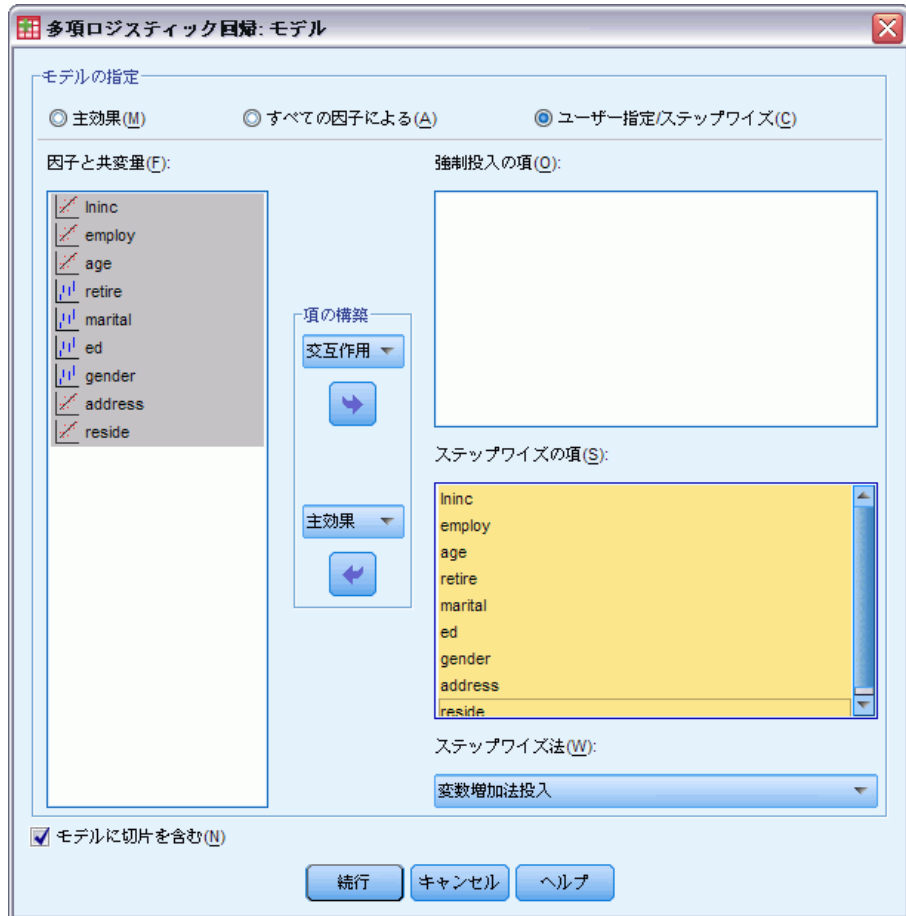
- ▶ 従属変数として「顧客カテゴリ」を選択します。
- ▶ 因子として、「婚姻状況」、「教育のレベル」、「退職」、および「性別」を選択します。
- ▶ 共変量として、「年齢」、「現住所の居住年数」、「勤続年数」、「世帯の同居人数」、「収入の対数」を選択します。
- ▶ 他の顧客を、基本サービスに加入している顧客と比較したい場合には、「顧客カテゴリ」を選択して、「参照カテゴリ」をクリックします。

図 5-31  
[参照カテゴリ] ダイアログ ボックス



- ▶ [最初のカテゴリ] を選択します。
- ▶ [続行] をクリックします。
- ▶ [多項ロジスティック回帰] ダイアログ ボックスで、[モデル] をクリックします。

図 5-32  
[モデル] ダイアログ ボックス



- ▶ [ユーザー指定/ステップワイズ] を選択します。
- ▶ [ステップワイズ項の構築] ドロップダウンから、[主効果] を選択します。
- ▶ [ステップワイズの項] として、「Ininc」から「同居人数」までを選択します。
- ▶ [続行] をクリックします。
- ▶ [多項ロジスティック回帰] ダイアログ ボックスで、[OK] をクリックします。

## ステップ要約

図 5-33  
ステップの要約

代入番号	モデル	アクション	効果	モデル当 てはめ基 準	効果選択検定		
				-2 対数 尤度	カイ 2 乗	自由度	有意確率
元のデータ	0	投入済み	切片	1.354E3	.		
	1	投入済み	ed	1.261E3	92.583	12	.000
	2	投入済み	employ	1.238E3	23.308	3	.000
	3	投入済み	marital	1.230E3	7.856	3	.049
1	0	投入済み	切片	2.763E3	.		
	1	投入済み	ed	2.600E3	162.621	12	.000
	2	投入済み	employ	2.562E3	38.359	3	.000
	3	投入済み	marital	2.548E3	13.608	3	.003
	4	投入済み	lninc	2.538E3	10.288	3	.016
	5	投入済み	address	2.528E3	9.602	3	.022
2	0	投入済み	切片	2.763E3	.		
	1	投入済み	ed	2.601E3	161.366	12	.000
	2	投入済み	employ	2.559E3	41.838	3	.000
	3	投入済み	lninc	2.545E3	14.814	3	.002
	4	投入済み	marital	2.530E3	14.983	3	.002
	5	投入済み	address	2.520E3	9.453	3	.024
3	0	投入済み	切片	2.763E3	.		
	1	投入済み	ed	2.605E3	157.180	12	.000
	2	投入済み	employ	2.563E3	42.768	3	.000
	3	投入済み	reside	2.550E3	12.854	3	.005
	4	投入済み	address	2.541E3	8.975	3	.030
4	0	投入済み	切片	2.763E3	.		
	1	投入済み	ed	2.605E3	158.011	12	.000
	2	投入済み	employ	2.561E3	43.690	3	.000
	3	投入済み	marital	2.546E3	14.489	3	.002
	4	投入済み	address	2.536E3	9.996	3	.019
	5	投入済み	lninc	2.528E3	8.268	3	.041
5	0	投入済み	切片	2.763E3	.		
	1	投入済み	ed	2.597E3	165.609	12	.000
	2	投入済み	employ	2.555E3	42.129	3	.000
	3	投入済み	marital	2.537E3	17.670	3	.001
	4	投入済み	address	2.525E3	11.693	3	.009

ステップワイズ手法: 前方エントリ

a. 投入に対するカイ 2 乗は尤度比検定に基づいています。

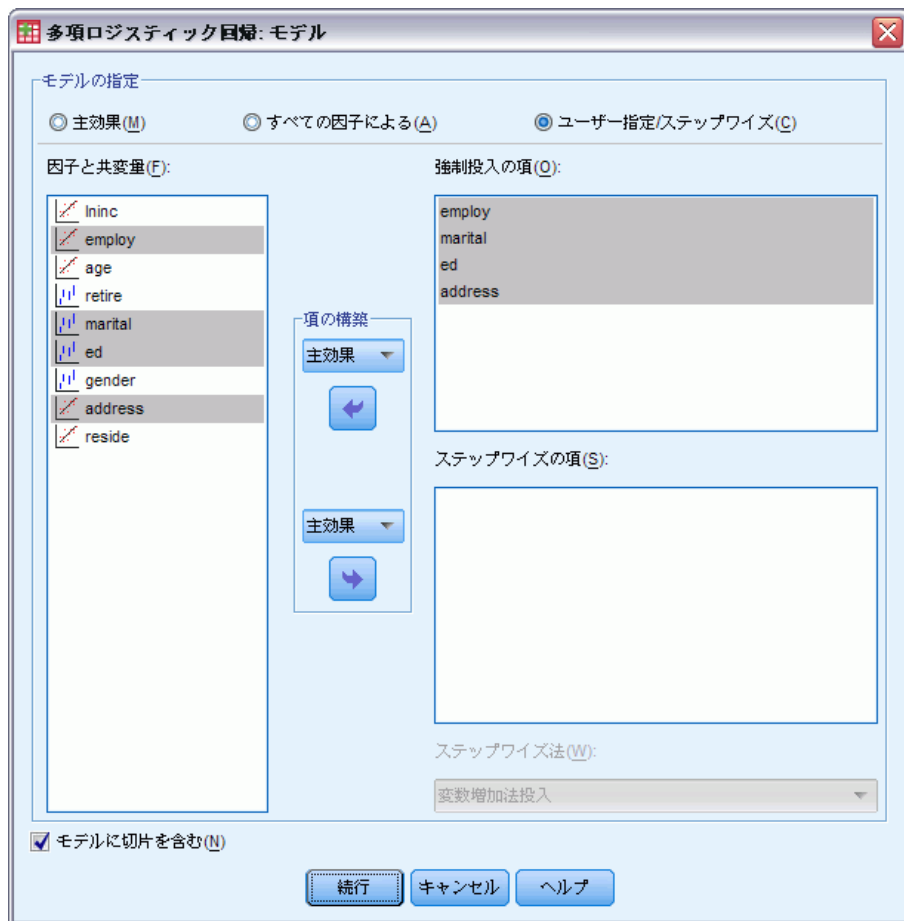
多項ロジスティック回帰は、回帰係数のプールをサポートします。ただし、出力の「all」テーブルは、代入と元のデータの結果を表すことに注意してください。これは、ファイルがImputation\_で分割されるためであり、したがって、分割変数を使用するすべてのテーブルは、分割ファイルグループをまとめて 1 つのテーブルに示します。

パラメータ推定値テーブルは、プールされた推定値を表示しないため、ステップの要約を確認します。モデル効果のステップワイズの選択を要求し、効果の同じセットはすべての代入で選択されていません。このた

め、プールを実行することはできません。ただし、これは、有益な情報を提供しています。「学歴（教育レベル）」、「雇用（勤続年数）」、「婚姻状況」、「居住年数」が、代入においてステップワイズの選択で頻繁に選択されているためです。この予測変数だけを使用して、他のモデルを適用しましょう。

## 予測変数のサブセットを使用してモデルを実行

図 5-34  
[モデル] ダイアログ



- ▶ [多項ロジスティック回帰] ダイアログ ボックスをもう一度開き、[モデル] をクリックします。
- ▶ [ステップワイズの項] リストから変数の選択を解除します。
- ▶ [強制投入の項の構築] ドロップダウンから、[主効果] を選択します。

- ▶ 強制投入の項として、「雇用」、「婚姻状況」、「学歴」、「居住年数」を選択します。
- ▶ [続行] をクリックします。
- ▶ [多項ロジスティック回帰] ダイアログ ボックスで、[OK] をクリックします。

## プールされたパラメータ推定値

このテーブルはかなり大きいですが、ピボットすることで、出力を様々な有益な視点で確認することができます。

図 5-35  
プールされたパラメータ推定値

代入番号	階層カテゴリ	Wald	自由度	有意確率	Exp (B)
元のデータ	基本サービス	5.177	1	.023	
		5.589	1	.018	.957
		3.530	1	.060	1.686
		24.807	1	.000	36.223
		13.752	1	.000	8.438
		4.016	1	.045	3.366
		.627	1	.428	1.597
		.461	1	.497	.988
E-サービス		1.954	1	.162	
		.350	1	.554	1.010
		.715	1	.398	.789
		3.757	1	.053	3.736
		.934	1	.334	1.609
		.521	1	.471	1.442
		.050	1	.822	.898
address		.016	.017	.922	1.016
追加サービス	切片	-19.177	.330	3372.746	1.000
	employ	.012	.016	.537	1.464
	marital=0	.027	.260	.010	.921

- ▶ テーブルをアクティブ化 (ダブルクリック) し、メニューから [ピボットトレイ] を選択します。

図 5-36  
プールされたパラメータ推定値

		Exp(B) の 95% 信頼						
		下限		上限				
代入番号	1			.923				
代入番号	2			.978			2	
代入番号	3			8.821		148		
代入番号	4			2.733		26		
代入番号	5			1.027		11		
代入番号	address			.501		5		
E-サービス	切片			.954		1		
E-サービス	employ			.976		1		
E-サービス	[marital=			.455		1		
E-サービス	[marital=			.985		14		
E-サービス	[ed=1]			.614		4		
E-サービス	[ed=2]	.475	.492	.934	1	.334	1.609	
E-サービス	[ed=3]	.366	.507	.521	1	.471	1.442	
E-サービス	[ed=4]	-.108	.481	.050	1	.822	.898	
E-サービス	[ed=5]	0 <sup>a</sup>			0			
E-サービス	address	.016	.017	.922	1	.337	1.016	
追加サービス	切片	-19.177	.330	3372.746	1	.000		

- ▶ 行から層に「反復回数」を移動します。
- ▶ [反復回数] ドロップダウン リストから [Pooled] を選択します。

図 5-37  
プールされたパラメータ推定値

顧客カテゴリ a,b,c,d,e,f	B	標準誤差	有意確率	Exp (B)	Exp(B) の 95% 信頼区間		欠損比の情報	相対増加分散	相対効率
					下限	上限			
基本サービス	切片	-1.123	.414	.007			.051	.052	.990
	employ	-.039	.012	.001	.962	.939	.985	.034	.993
	[marital=0]	.681	.204	.001	1.976	1.322	2.954	.121	.976
	[marital=1]	0 <sup>a</sup>							
	[ed=1]	3.541	.528	.000	34.489	12.249	97.110	.065	.987
	[ed=2]	1.804	.436	.000	6.074	2.582	14.291	.071	.986
	[ed=3]	1.331	.435	.002	3.786	1.613	8.885	.021	.996
	[ed=4]	.565	.429	.188	1.759	.758	4.082	.048	.990
	[ed=5]	0 <sup>a</sup>							
address	-.018	.012	.131	.982	.960	1.005	.029	.994	
E-サービス	切片	-.528	.317	.095			.015	.015	.997
	employ	-.011	.011	.341	.989	.967	1.012	.033	.994
	[marital=0]	.120	.208	.565	1.127	.748	1.700	.161	.969
	[marital=1]	0 <sup>a</sup>							
	[ed=1]	1.437	.466	.002	4.208	1.687	10.496	.015	.997
	[ed=2]	.418	.348	.231	1.518	.767	3.006	.023	.995
	[ed=3]	.508	.348	.145	1.661	.840	3.286	.005	.999
	[ed=4]	-.072	.330	.827	.930	.487	1.776	.010	.998
	[ed=5]	0 <sup>a</sup>							
address	.016	.011	.146	1.016	.994	1.039	.039	.992	
追加サービス	切片	-2.326	.552	.000			.005	.005	.999
	employ	.011	.011	.295	1.012	.990	1.033	.050	.990
	[marital=0]	.335	.200	.094	1.398	.944	2.070	.092	.982
	[marital=1]	0 <sup>a</sup>							
	[ed=1]	4.104	.632	.000	60.607	17.557	209.223	.013	.997
	[ed=2]	2.613	.561	.000	13.646	4.542	40.995	.005	.999
	[ed=3]	2.078	.570	.000	7.985	2.614	24.391	.002	1.000
	[ed=4]	1.083	.569	.057	2.955	.969	9.008	.004	.999
	[ed=5]	0 <sup>a</sup>							
address	.002	.011	.892	1.002	.980	1.024	.054	.989	

このビューは、プールされた結果のすべての統計量を示します。欠損値のないデータセットのテーブルと同様に、この係数を使用したり、解釈することができます。

パラメータ推定値テーブルは、各予測変数の効果をまとめたものです。この係数と標準誤差の比を 2 乗すると、Wald 統計量に等しくなります。Wald 統計量の有意水準が小さい (0.05 未満) 場合、そのパラメータは 0 ではありません。

- 有意な負の係数をもつパラメータは、参照カテゴリに対する応答カテゴリの尤度を減少させます。
- 正の係数を持つパラメータで、その応答カテゴリの尤度が大きくなります。
- 定数項がある場合、各因子の最後のカテゴリと関連するパラメータは冗長です。

これらは、テーブルに追加される 3 つの列で、プールされた出力の詳細な情報を提供します。「欠損情報の割合」は、「完全な」情報に対する欠損情報の割合推定値で、反応がない場合は、「分散の相対増加」に基づきます。これは、代入間と回帰係数の代入分散の平均の (調整された) 割合です。「相対効率」は、この推定値と代入の無限数で計算された (理論上の)



推定値の比較です。相対効率率は、欠損情報の割合とプールされた結果を得るために使用された代入回数によって決定されます。欠損情報の割合が大きい場合、より多くの代入で、相対効率を 1 に近づけ、プールされた推定値を理想的な推定値に近づける必要があります。

図 5-38  
プールされたパラメータ推定値

変数	層	列	推定値	標準誤差	Wald	自由度	有意確率	Exp(B)	Exp(B)の95%信頼区間 下限	Exp(B)の95%信頼区間 上限	
顧客力 基本サービス	[ed=2]		2								
	[ed=3]		1								
	[ed=4]										
	[ed=5]										
	address										
	E-サービス	切片									
		employ									
[marital=0]											
[marital=1]											
[ed=1]			1								
[ed=2]				.475	.403	.372	.426	.496	.391	.418	
[ed=3]				.366	.529	.497	.515	.524	.472	.508	
[ed=4]			-.108	-.079	-.071	-.050	-.044	-.118	-.072		
[ed=5]			0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>		
address			.016	.014	.016	.015	.018	.019	.016		
追加サービス	切片		-19.177	-2.294	-2.286	-2.336	-2.360	-2.357	-2.326		
	employ		.012	.011	.013	.013	.013	.008	.011		

- ▶ もう一度テーブルをアクティブ化（ダブルクリック）し、メニューから [ピボットトレイ] を選択します。
- ▶ 層から列に「反復回数」を移動します。
- ▶ 列から層に「統計量」を移動します。
- ▶ ドロップダウン リストから、[B] を選択します。

図 5-39  
プールされたパラメータ推定値、列の反復数と層の統計量

顧客カテゴリ   a,b,c,d,e,f	代入番号						プールされた	
	元のデータ	1	2	3	4	5		
基本サービス	切片	-1.266	-1.148	-1.041	-1.027	-1.188	-1.210	-1.123
	employ	-.044	-.036	-.038	-.039	-.040	-.041	-.039
	[marital=0]	.522	.682	.672	.601	.673	.778	.681
	[marital=1]	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
	[ed=1]	3.590	3.628	3.442	3.413	3.524	3.696	3.541
	[ed=2]	2.133	1.823	1.701	1.691	1.905	1.901	1.804
	[ed=3]	1.214	1.352	1.255	1.293	1.358	1.398	1.331
	[ed=4]	.468	.635	.465	.479	.629	.616	.565
	[ed=5]	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
	address	-.012	-.020	-.019	-.016	-.016	-.018	-.018
E-サービス	切片	-.629	-.517	-.536	-.479	-.575	-.535	-.528
	employ	.010	-.011	-.009	-.010	-.011	-.014	-.011
	[marital=0]	-.237	.155	.118	.027	.079	.221	.120
	[marital=1]	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
	[ed=1]	1.318	1.508	1.407	1.369	1.459	1.442	1.437
	[ed=2]	.475	.403	.372	.426	.496	.391	.418
	[ed=3]	.366	.529	.497	.515	.524	.472	.508
	[ed=4]	-.108	-.079	-.071	-.050	-.044	-.118	-.072
	[ed=5]	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
	address	.016	.014	.016	.015	.018	.019	.016
追加サービス	切片	-19.177	-2.294	-2.286	-2.336	-2.360	-2.357	-2.326
	employ	.012	.011	.013	.013	.013	.008	.011
	[marital=0]	-.027	.303	.321	.303	.318	.431	.335
	[marital=1]	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
	[ed=1]	21.124	4.177	4.073	4.041	4.060	4.171	4.104
	[ed=2]	19.486	2.593	2.594	2.580	2.672	2.629	2.613
	[ed=3]	18.976	2.114	2.060	2.087	2.072	2.055	2.078
	[ed=4]	18.374	1.100	1.043	1.088	1.128	1.058	1.083
	[ed=5]	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
	address	.011	.001	-.002	.003	.002	.004	.002

このテーブルビューは、代入から代入への回帰係数推定値の変動を視覚的に確認するために、代入の値同士を比較したり、元のデータと比較するのに役立ちます。特に、統計量を層から、Std. Error に切り替えることで、多重代入によって、係数推定値とリストごとの削除（元のデータ）との比較において、変動がどれくらい減少したかを確認することができます。

図 5-40  
警告

次の変数: ln収入, 年齢, 退職, 性別, 住所 は、部分母集団の定義にのみ使用され、モデルの構築には使用されません。  
ヘッセ行列に予期しない特異性が見つかりました。これは、いくつかの予測変数を除外するか、いくつかのカテゴリを結合する必要があることを示しています。  
NOMREG 手続きは、上記の警告が表示されても継続します。その後表示される結果は、最後の反復に基づいています。モデルの適合度が有効かどうかは不明です。

ただし、この例では、元のデータセットが実際にエラーを発生させます。このエラーは、テーブルの元のデータ列の「プラス サービス」定数項と「学歴（教育レベル）」の非冗長レベルの非常に大きなパラメータ推定値の原因です。

## [要約]

多重代入手順を使用して、欠損値のパターンを分析して、単純なリストごとの削除を使用すると、多くの情報が失われる可能性があることが分かりました。最初に多重代入を自動実行した後で、代入値を妥当な範囲内に収めるには、制約条件が必要であることが分かりました。制約条件を設けて実行することで、適切な値が作成されます。FCS 方法が収束しないことの明白は証拠はありませんでした。多重代入値で「完成した」データセットを使用することで、多項ロジスティック回帰をデータに適用して、プールされた回帰推定値を入手し、さらに、元のデータでリストごとの削除を使用したのでは実際にできなかった、最終モデルが適用できることが分かりました。

# サンプル ファイル

製品とともにインストールされるサンプル ファイルは、インストールディレクトリの Samples サブディレクトリにあります。[サンプル] サブディレクトリ内に次の各言語の別のフォルダがあります。英語、フランス語、ドイツ語、イタリア語、日本語、韓国語、ポーランド語、ロシア語、簡体字中国語、スペイン語、そして繁体中国語です。

すべてのサンプル ファイルが、すべての言語で使用できるわけではありません。サンプル ファイルがある言語で使用できない場合、その言語のフォルダには、サンプル ファイルの英語バージョンが含まれています。

## 説明

以下は、このドキュメントのさまざまな例で使用されているサンプル ファイルの簡単な説明です。

- **accidents.sav**。与えられた地域での自動車事故の危険因子を年齢および性別ごとに調べている保険会社に関する架空のデータ ファイルです。各ケースが、年齢カテゴリと性別のクロス分類に対応します。
- **adl.sav**。脳卒中患者に提案される治療の効果を特定するための取り組みに関する架空のデータ ファイルです。医師団は、女性の脳卒中患者たちを、2 つのグループのいずれかにランダムに割り当てました。一方のグループは標準的な理学療法を受け、もう一方のグループは感情面の治療も追加で受けました。治療の 3 か月後に、各患者が日常生活の一般的な行動をどの程度とることができるかを、順序変数として得点付けしました。
- **advert.sav**。広告費とその売上成果の関係を調べるための小売業者の取り組みに関する架空のデータ ファイルです。この小売業者は、そのために、過去の売上と、それに関する広告費のデータを収集しました。
- **aflatoxin.sav**。収穫物によって濃度が大きく異なる毒物であるアフラトキシンを、トウモロコシの収穫物に関して検定することに関する架空のデータ ファイルです。ある穀物加工業者は、8 つそれぞれの収穫物から 16 のサンプルを受け取って、10 億分の 1 単位でアフラトキシン レベルを測定しました。
- **anorectic.sav**。拒食行動または過食行動の標準的な症状の特定を目指して、調査員 が、摂食障害を持つ大人 55 人の調査を行いました。各患者が 4 年間で 4 回診察を受けたので、観測値は合計で 220 になりました。観測値ごとに、16 種類の症状に関して患者の得点が記録

されました。患者 71 (2 回目)、患者 76 (2 回目)、患者 47 (3 回目) の症状の得点が見つからなかったので、残っている 217 回分の観測値が有効です。

- **bankloan.sav.** 債務不履行率を低減させるための銀行の取り組みに関する架空のデータ ファイルです。このファイルには、過去の顧客および見込み客 850 人に関する財務情報と人口統計情報が含まれています。最初の 700 ケースは、以前に貸付を行った顧客です。残りの 150 ケースは見込み顧客で、これらの顧客に関して銀行は信用リスクの良し悪しを分類する必要があります。
- **bankloan\_binning.sav.** 過去の顧客 5,000 人に関する財務情報と人口統計情報を含む架空のデータ ファイルです。
- **behavior.sav.** 52 人の学生に 15 の状況と 15 の行動の組み合わせについて、0 = 「非常に適切」から 9 = 「非常に不適切」までの 10 段階でランク付けするよう依頼した研究があります。個人間の平均を取ったため、値は非類似度としてみなされます。
- **behavior\_ini.sav.** このデータ ファイルには、behavior.sav の 2 次元の解の初期配置が含まれています。
- **brakes.sav.** 高性能自動車のディスク ブレーキを生産している工場での品質管理に関する架空のデータ ファイルです。このデータ ファイルには、8 台の機械で生産した 16 個のディスクの直径測定値が含まれています。ブレーキの目標の直径は 322 ミリメートルです。
- **breakfast.sav.** 21 人の Wharton School MBA の学生およびその配偶者に、15 種類の朝食を好みの順に (1 = 「最も好き」から 15 = 「最も嫌い」まで) ランク付けするよう依頼した研究があります。調査対象者の嗜好は、「すべて」から「スナックとドリンクのみ」まで、6 つの異なるシナリオに基づいて記録されました。
- **breakfast-overall.sav.** このデータ ファイルには、最初のシナリオ (「すべて」) のみの朝食の好みが含まれています。
- **broadband\_1.sav.** 全国規模のブロードバンド サービスの地域ごとの契約者数を含む架空のデータ ファイルです。このデータ ファイルには、85 地域の月々の契約者数が 4 年間分含まれています。
- **broadband\_2.sav.** このデータ ファイルは broadband\_1.sav と同じですが、データが 3 か月分追加されています。
- **car\_insurance\_claims.sav.** 他の場所 で表示および分析される、自動車の損害請求に関するデータセットです。逆リンク関数を使用して従属変数の平均値を保険契約者の年齢、車種、製造年の線型結合と関連付けることにより、平均請求数はガンマ分布としてモデリングできます。申請された請求の数は、尺度重み付けとして使用できます。
- **car\_sales.sav.** このデータ ファイルには、自動車のさまざまな車種やモデルの架空の売上推定値、定価、仕様が含まれています。定価と仕様はそれぞれ、edmunds.com と製造元のサイトから入手しました。

- **car\_sales\_upprepared.sav.** 変換したバージョンのフィールドを含まない car\_sales.sav の修正したバージョンです。
- **carpet.sav.** 一般的な例 としては、新しいカーペット専用洗剤を市販することに関心のある企業が消費者の嗜好に関する 5 種類の因子（パッケージのデザイン、ブランド名、価格、サービスシール、料金の払い戻し）の影響について調べたい場合があります。パッケージのデザインには、3 つの因子レベルがあります。それぞれ塗布用ブラシの位置が異なります。また、3 つのブランド名（K2R、Glory、および Bissell）、3 つの価格水準があり、最後の 2 つの因子のそれぞれに対しては 2 つのレベル（「なし」または「あり」）があります。10 人の消費者が、これらの因子により定義された 22 個のプロファイルに順位を付けます。変数「嗜好」には、各プロファイルの平均順位の序列が含まれています。順位が低いほど、嗜好度は高くなります。この変数には、各プロファイルの嗜好測定値がすべて反映されます。
- **carpet\_prefs.sav.** このデータ ファイルは carpet.sav と同じ例に基づいていますが、10 人の消費者それぞれから収集した実際のランキングが含まれています。消費者は、22 種類の製品プロファイルを、一番好きなものから一番嫌いなものまで順位付けすることを依頼されています。変数 PREF1 から PREF22 には、carpet\_plan.sav で定義されている、関連するプロファイルの ID が含まれています。
- **catalog.sav.** このデータ ファイルには、あるカタログ会社が販売した 3 つの製品の、架空の月間売上高が含まれています。5 つの予測変数のデータも含まれています。
- **catalog\_seasfac.sav.** このデータ ファイルは catalog.sav と同じですが、季節性の分解手続きとそれに付随する日付変数から計算した一連の季節因子が追加されています。
- **cellular.sav.** 解約率を削減するための携帯電話会社の取り組みに関する架空のデータ ファイルです。解約の傾向スコアは、0 ~ 100 の範囲でアカウントに適用されます。スコアリングが 50 以上のアカウントはプロバイダの変更を考えている場合があります。
- **ceramics.sav.** 新しい上質の合金に標準的な合金より高い耐熱性があるかどうかを特定するための、ある製造業者の取り組みに関する架空のデータ ファイルです。各ケースが 1 つの合金の別々のテストを表し、軸受けの耐熱温度が記録されます。
- **cereal.sav.** 880 人を対象に、朝食の好みについて、年齢、性別、婚姻状況、ライフスタイルが活動的かどうか（週 2 回以上運動するか）を含めて調査した、架空のデータ ファイルです。各ケースが別々の回答者を表します。
- **clothing\_defects.sav.** ある衣料品工場での品質管理工程に関する架空のデータ ファイルです。工場で生産される各ロットから、調査員が衣料品のサンプルを取り出し、不良品の数を数えます。

- **coffee.sav.** このデータ ファイルは、6 つのアイスコーヒー ブランド について受けた印象に関連しています。回答者は、アイス コーヒーに対する 23 の各印象属性に対して、その属性が言い表していると思われるすべてのブランドを選択しました。機密保持のため、6 つのブランドを AA、BB、CC、DD、EE、および FF で表しています。
- **contacts.sav.** 企業のコンピュータ営業グループの担当者リストに関する架空のデータ ファイルです。各担当者は、所属する会社の部門および会社のランクによって分類されています。また、最新の販売金額、最後の販売以降の経過時間、担当者の会社の規模も記録されています。
- **creditpromo.sav.** 最近のクレジット カード プロモーションの有効性を評価するための、あるデパートの取り組みに関する架空のデータ ファイルです。このために、500 人のカード所有者がランダムに選択されました。そのうち半分には、今後 3 か月間の買い物に関して利率を下げることをプロモーションする広告を送付しました。残り半分には、通常どおりの定期的な広告を送付しました。
- **customer\_dbase.sav.** 自社のデータ ウェアハウスにある情報を使用して、反応がありそうな顧客に対して特典を提供するための、ある会社の取り組みに関する架空のデータ ファイルです。顧客ベースのサブセットをランダムに選択して特典を提供し、顧客の反応が記録されています。
- **customer\_information.sav.** 名前や住所など、顧客の連絡先情報を含む架空のデータ ファイルです。
- **customer\_subset.sav.** customer\_dbase.sav の 80 件のケースのサブセット。
- **debate.sav.** 政治討論の出席者に対して行った調査の、討論の前後それぞれの回答に関する架空のデータ ファイルです。各ケースが別々の回答者に対応します。
- **debate\_aggregate.sav.** debate.sav 内の回答を集計する、架空のデータ ファイルです。各ケースが、討論前後の好みのクロス分類に対応しています。
- **demo.sav.** 月々の特典を送付することを目的とした、購入顧客のデータベースに関する架空のデータ ファイルです。顧客が特典に反応したかどうか、さまざまな人口統計情報と共に記録されています。
- **demo\_cs\_1.sav.** 調査情報のデータベースをコンパイルするための、ある会社の取り組みの最初のステップに関する架空のデータ ファイルです。各ケースが別々の都市に対応し、地域、地方、地区、および都市の ID が記録されています。
- **demo\_cs\_2.sav.** 調査情報のデータベースをコンパイルするための、ある会社の取り組みの第 2 のステップに関する架空のデータ ファイルです。各ケースが、最初のステップで選択した都市の別々の世帯単位に対応し、地域、地方、地区、都市、区画、および単位の ID が記録されます。計画の最初の 2 つの段階からの抽出情報も含まれています。

- **demo\_cs.sav**。コンプレックス サンプル計画を使用して収集された調査情報を含む架空のデータ ファイルです。各ケースが別々の世帯単位に対応し、さまざまな人口統計情報および抽出情報が記録されています。
- **dmdata.sav**。これは、ダイレクト マーケティング企業の人口統計情報および購入情報を含む架空のデータです。dmdata2.sav には、テストメールを受け取った連絡先のサブセットの情報を含み、dmdata3.sav には、テストメールを受け取らなかった残りの連絡先に関する情報を含みます。
- **dietstudy.sav**。この架空のデータ ファイルには、“Stillman diet” の研究結果が含まれています。各ケースが別々の被験者に対応し、被験者のダイエット前後の体重（ポンド単位）と、トリグルセリド レベル（mg/100 ml 単位）が記録されています。
- **dvdplayer.sav**。新しい DVD プレーヤーの開発に関する架空のデータ ファイルです。プロトタイプを使用して、マーケティング チームはフォーカス グループ データを収集しました。各ケースが別々の調査対象ユーザーに対応し、ユーザーの人口統計情報と、プロトタイプに関する質問への回答が記録されています。
- **german\_credit.sav**。このデータ ファイルは、カリフォルニア大学アーバイン校の Repository of Machine Learning Databases にある “German credit” データセットから取ったものです。
- **grocery\_1month.sav**。この架空のデータ ファイルは、grocery\_coupons.sav データ ファイルの週ごとの購入を「ロールアップ」して、各ケースが別々の顧客に対応するようにしたものです。その結果、週ごとに変っていた変数の一部が表示されなくなり、買物の総額が、調査を行った 4 週間の買物額の合計になっています。
- **grocery\_coupons.sav**。顧客の購買習慣に関心を持っている食料雑貨店チェーンが収集した調査データを含む架空のデータ ファイルです。各顧客を 4 週間に渡って追跡し、各ケースが別々の顧客の週に対応しています。その週に食料品に費やした金額も含め、顧客がいつどこで買物をするかに関する情報が記録されています。
- **guttman.sav**。Bell は、予想される社会グループを示す表を作成しました。Guttman は、この表の一部を使用しました。この表では、社会相互作用、グループへの帰属感、メンバとの物理的な近接性、関係の形式化などを表す 5 個の変数が、理論上の 7 つの社会グループと交差しています。このグループには、観衆（例、フットボールの試合の観戦者）、視聴者（例、映画館または授業の参加者）、公衆（例、新聞やテレビの視聴者）、暴徒（観衆に似ているが、より強い相互作用がある）、第一次集団（親密な関係）、第二次集団（自発的な集団）、および近代コミュニティ（物理的により密接した近接性と特化されたサービスの必要性によるゆるい同盟関係）があります。



- **health\_funding.sav**。医療用資金（人口 100 人あたりの金額）、罹患率（人口 10,000 人あたりの人数）、医療サービス機関への訪問率（人口 10,000 人あたりの人数）のデータを含む、架空のデータ ファイルです。各ケースが別々の都市を表します。
- **hivassay.sav**。HIV 感染を発見する迅速な分析方法を開発するための、ある製薬研究所の取り組みに関する架空のデータ ファイルです。分析の結果は、8 段階の濃さの赤で表現され、色が濃いほど感染の可能性が高くなります。研究所では 2,000 件の血液サンプルに関して試験を行い、その半数が HIV に感染しており、半分は感染していませんでした。
- **hourlywagedata.sav**。管理職から現場担当まで、またさまざまな経験レベルの看護師の時給に関する架空のデータ ファイルです。
- **insurance\_claims.sav**。不正請求の恐れがある、疑いを区別するためにモデルを作成する必要がある保険会社の仮説データ ファイルです。各ケースがそれぞれの請求を表します。
- **insure.sav**。10 年満期の生命保険契約に対し、顧客が請求を行うかどうかを示す危険因子を調査している保険会社に関する架空のデータ ファイルです。データ ファイルの各ケースは、年齢と性別が一致する、請求を行った契約と行わなかった契約のペアを表します。
- **judges.sav**。訓練を受けた審判（および 1 人のファン）が 300 件の体操の演技に対して付けた得点に関する架空のデータ ファイルです。各行が別々の演技を表し、審判たちは同じ演技を見ました。
- **kinship\_dat.sav**。Rosenberg と Kim は、15 種類の親族関係用語（祖父、祖母、父、母、叔父、叔母、兄弟、姉妹、いとこ、息子、娘、甥、姪、孫息子、孫娘）の分析を行いました。Rosenberg と Kim は、大学生の 4 つのグループ（女性 2 組、男性 2 組）に、類似性に基づいて上記の用語を並べ替えるよう依頼しました。2 つのグループ（女性 1 組、男性 1 組）には、1 回目と違う条件に基づいて、2 回目の並べ替えをするように頼みました。このようにして、合計で 6 つの「ソース」が取得できました。各ソースは、15 × 15 の近接行列に対応します。この近接行列のセルの数は、ソースの人数から、ソース内でオブジェクトを分割した回数を引いたものです。
- **kinship\_ini.sav**。このデータ ファイルには、kinship\_dat.sav の 3 次元の解の初期布置が含まれています。
- **kinship\_var.sav**。このデータ ファイルには、kinship\_dat.sav の解の次元の解釈に使用できる独立変数である性別、世代、および(ation), and 親等が含まれています。特に、解の空間をこれらの変数の線型結合に制限するために使用できます。
- **marketvalues.sav**。1999 ~ 2000 年の間の、イリノイ州アルゴンキンの新興住宅地での住宅売上に関するデータ ファイルです。これらの売上は、公開レコードの問題となります。

- **nhis2000\_subset.sav**。National Health Interview Survey (NHIS) は、米国民を対象とした人口ベースの大規模な調査です。全国の代表的な世帯サンプルについて対面式で調査が行われます。各世帯のメンバーに関して、人口統計情報、健康に関する行動および状態の観測値が得られます。このデータ ファイルには、2000 年の調査から得られた情報のサブセットが含まれています。National Center for Health Statistics。National Health Interview Survey, 2000。一般使用データおよびドキュメント。[ftp://ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Datasets/NHIS/2000/](ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/)。2003 年にアクセス。
- **ozone.sav**。データには、残りの変数からオゾン濃度を予測するための、6 個の気象変数に対する 330 個の観測値が含まれています。それまでの研究者、が、他の研究者と共に、これらの変数間に非線型性を確認しています。この場合、標準的な回帰アプローチは使用できません。
- **pain\_medication.sav**。この架空のデータ ファイルには、慢性関節炎を治療する抗炎症薬の臨床試験の結果が含まれています。特に興味深いことは、薬の効果が出るまでの時間と、既存の薬剤との比較です。
- **patient\_los.sav**。この架空のデータ ファイルには、心筋梗塞 (MI、または「心臓発作」) の疑いで入院した患者の治療記録が含まれています。各ケースが別々の患者に対応し、入院に関連する多くの変数が記録されています。
- **patlos\_sample.sav**。この架空のデータ ファイルには、心筋梗塞 (MI、または「心臓発作」) の治療中に血栓溶解剤を投薬された患者のサンプルの治療記録が含まれています。各ケースが別々の患者に対応し、入院に関連する多くの変数が記録されています。
- **poll\_cs.sav**。市民の法案支持率を議会開会前に特定するための、世論調査員の取り組みに関する架空のデータ ファイルです。各ケースは登録有権者に対応しています。ケースごとに、有権者が居住している郡、町、区域が記録されています。
- **poll\_cs\_sample.sav**。この架空のデータ ファイルには、poll\_cs.sav の有権者のサンプルが含まれています。サンプルは、poll\_csplan 計画ファイルで指定されている計画に従って抽出され、このデータ ファイルには包含確率およびサンプル重み付けが記録されています。ただし、抽出計画では確率比例 (PPS) 法を使用するため、結合選択確率を含むファイル (poll\_jointprob.sav) もあります。サンプル抽出後、有権者の人口統計および法案に関する意見に対応する追加の変数が収集され、データ ファイルに追加されました。
- **property\_assess.sav**。限られたリソースで資産価値評価を最新に保つための、郡の評価担当者の取り組みに関する架空のデータ ファイルです。各ケースは、前年に郡内で売却された資産に対応します。データ ファイル内の各ケースでは、資産が存在する町、最後に訪問した評価担当者、その評価からの経過時間、当時行われた評価、および資産の売却価値が記録されています。

- **property\_assess\_cs.sav.** 限られたリソースで資産価値評価を最新に保つための、州の評価担当者の取り組みに関する架空のデータ ファイルです。各ケースは州内の資産に対応します。データ ファイル内の各ケースでは、資産が存在する郡、町、および区域、最後の評価からの経過時間、および当時行われた評価が記録されています。
- **property\_assess\_cs\_sample.sav.** この架空のデータ ファイルには、property\_assess\_cs.sav の資産のサンプルが含まれています。サンプルは、property\_assess\_csplan 計画ファイルで指定されている計画に従って抽出され、このデータ ファイルには包含確率およびサンプル重み付けが記録されています。サンプル抽出後、現在の価値変数が収集され、データ ファイルに追加されました。
- **recidivism.sav.** 管轄地域での累犯率を把握するための、政府の法執行機関の取り組みに関する架空のデータ ファイルです。各ケースは元犯罪者に対応し、人口統計情報、最初の犯罪の詳細、初犯から 2 年以内の場合は 2 回目の逮捕までの期間が記録されています。
- **recidivism\_cs\_sample.sav.** 管轄地域での累犯率を把握するための、政府の法執行機関の取り組みに関する架空のデータ ファイルです。各ケースは 2003 年の 7 月に最初の逮捕から釈放された元犯罪者に対応し、人口統計情報、最初の犯罪の詳細、2006 年 7 月までの 2 回目の逮捕のデータが記録されています。犯罪者は recidivism\_cs.csplan で指定された抽出計画に従って抽出された部門から選択されます。調査では確率比例 (PPS) 法を採用したため、結合選択確率を保持したファイル (recidivism\_cs\_jointprob.sav) も用意されています。
- **rfm\_transactions.sav.** 購入日、購入品目、各取引のマネタリー量など、購買取引データを含む架空のデータ ファイルです。
- **salesperformance.sav.** 2 つの新しい販売トレーニング コースの評価に関する架空のデータ ファイルです。60 人の従業員が 3 つのグループに分けられ、全員が標準のトレーニングを受けます。さらに、グループ 2 は技術トレーニングを、グループ 3 は実践的なチュートリアルを受けます。トレーニング コースの最後に各従業員がテストを受け、得点が記録されました。データ ファイルの各ケースは別々の訓練生を表し、割り当てられたグループと、テストの得点が記録されています。
- **satisf.sav.** ある小売業者が 4 箇所の店舗で行った満足度調査に関する架空のデータ ファイルです。合計で 582 人の顧客を調査し、各ケースは 1 人の顧客からの回答を表します。
- **screws.sav.** このデータ ファイルには、ねじ、ボルト、ナット、鋸 (びょう) の特性に関する情報が含まれています。
- **shampoo\_ph.sav.** あるヘアケア製品工場での品質管理に関する架空のデータ ファイルです。定期的に、6 つの異なる製品が測定され、pH が記録されます。目標範囲は 4.5 ~ 5.5 です。

- **ships.sav.** 他の場所 で表示および分析される、波による貨物船への損害に関するデータセットです。件数は、船舶の種類、建造期間、およびサービス期間によって、ポワゾン率で発生するものとしてモデリングできます。因子のクロス分類によって形成されたテーブルの各セルのサービスマ月の集計によって、危険にさらされる確率の値が得られます。
- **site.sav.** 業務拡大に向けて新たな用地を選択するための、ある会社の取り組みに関する架空のデータ ファイルです。2 人のコンサルタントを雇って、用地を別々に評価させました。広範囲のレポートに加えて、各用地を「良い」、「普通」、「悪い」のいずれかで集計しました。
- **smokers.sav.** このデータ ファイルは、1998 年の National Household Survey of Drug Abuse から抜粋したものであり、アメリカの世帯の確率サンプルです。(<http://dx.doi.org/10.3886/ICPSR02934>) したがって、このデータ ファイルを分析する場合は、まず人口の傾向を反映させてデータを重み付けする必要があります。
- **stocks.sav** このデータ ファイルには、1 年あたりの在庫価格、量が含まれています。
- **stroke\_clean.sav.** この架空のデータ ファイルには、[データの準備] オプションの手続きを使用して整理した後の、医療データベースの状態が含まれています。
- **stroke\_invalid.sav.** この架空のデータ ファイルには、医療データベースの初期状態が含まれており、データ入力にいくつかエラーがあります。
- **stroke\_survival.** この架空のデータ ファイルは、虚血性脳卒中で数回の困難に直面した後リハビリ プログラムを終えた患者の生存時間に関するものです。脳卒中後、心筋梗塞の発生、虚血性脳卒中、または出血性脳卒中が注意され、イベントの時間が記録されます。脳卒中後に実施されたリハビリ プログラムの最後まで生存した患者のみが含まれるため、サンプルは左側が切り捨てられます。
- **stroke\_valid.sav.** この架空のデータ ファイルには、[データの検証] 手続きを使用して確認した後の、医療データベースの状態が含まれています。異常である可能性のあるケースが含まれています。
- **survey\_sample.sav.** このデータ ファイルには、人口統計データおよびさまざまな態度指標などの調査データが含まれています。これは「1998 NORC General Social Survey」の変数のサブセットに基づいていますが、いくつかのデータ値が変更され、追加の架空変数がデモの目的で追加されています。
- **telco.sav.** 顧客ベースにおける解約率を削減するための電気通信会社の取り組みに関する架空のデータ ファイルです。各ケースが別々の顧客に対応し、人口統計やサービス利用状況などのさまざまな情報が記録されています。
- **telco\_extra.sav.** このデータ ファイルは telco.sav データ ファイルに似ていますが、「期間」および対数変換された顧客支出の属性が削除され、標準化された対数変換顧客支出の変数に置き換えられています。

- **telco\_missing.sav**。このデータ ファイルは telco.sav データ ファイルのサブセットですが、一部の人口統計データ値が欠損値に置き換えられています。
- **testmarket.sav**。この架空のデータ ファイルは、新しいメニューを追加しようというファースト フード チェーンの計画に関連しています。新製品をプロモーションするためのキャンペーンには 3 つの候補があるため、新メニューはいくつかのランダムに選択した市場にある場所で紹介されます。場所ごとに別々のプロモーションを使用し、最初の 4 週間の新メニューの週間売上高が記録されます。各ケースが場所と週に対応します。
- **testmarket\_1month.sav**。この架空のデータ ファイルは、testmarket.sav データ ファイルの週ごとの売上を「ロールアップ」して、各ケースが別々の場所に対応するようにしたものです。その結果、週ごとに変っていた変数の一部が表示されなくなり、売上高が、調査を行った 4 週間の売上高の合計になっています。
- **tree\_car.sav**。これは、人口統計および自動車購入価格のデータを含む架空のデータ ファイルです。
- **tree\_credit.sav**。これは、人口統計および銀行ローン履歴のデータを含む架空のデータ ファイルです。
- **tree\_missing\_data.sav**。これは、人口統計および銀行ローン履歴のデータと、多数の欠損値を含む架空のデータ ファイルです。
- **tree\_score\_car.sav**。これは、人口統計および自動車購入価格のデータを含む架空のデータ ファイルです。
- **tree\_textdata.sav**。尺度および値ラベルを割り当てる前の、変数のデフォルトの状態を示すことを主な目的とする、変数を 2 つだけ含む単純なデータ ファイルです。
- **tv-survey.sav**。テレビ スタジオで実施された、ヒットした番組の放送期間を延長するかどうかを検討する調査に関する架空のデータ ファイルです。906 人の回答者に、さまざまな条件下でこの番組を視聴するかどうかを質問しました。各行は別々の回答者を表し、各列は別々の条件を表します。
- **ulcer\_recurrence.sav**。このファイルには、潰瘍の再発を防ぐための 2 つの治療の有効性を比較するように計画された調査の情報の一部が含まれています。これは区間調査の良い例であり、他の場所 で表示および分析されています。
- **ulcer\_recurrence\_recoded.sav**。このファイルでは、ulcer\_recurrence.sav の情報が、単に調査終了時のイベント確率ではなく調査の区間ごとのイベント確率をモデリングできるように再編成されています。これは他の場所 で表示および分析されています。
- **verd1985.sav**。このデータ ファイルは調査に関連しています。8 つの変数に対する 15 人の被験者の回答を記録しました。対象となる変数が 3 つのグループに分類されます。グループ 1 には「年齢」と「婚姻」、

グループ 2 には「ペット」と「新聞」、グループ 3 には「音楽」と「居住地域」がそれぞれ含まれます。「ペット」は多重名義として尺度化され、「年齢」は順序として尺度化されます。また、その他のすべての変数は単一名義として尺度化されます。

- **virus.sav**。自社のネットワーク上のウィルスの影響を特定するための、インターネット サービス プロバイダ (ISP) の取り組みに関する架空のデータ ファイルです。この ISP は、ネットワーク上の感染した E メール トラフィックの (およその) パーセンテージを、発見の瞬間から脅威が阻止されるまで追跡しました。
- **wheeze\_steubenville.sav**。これは、子供 に対する大気汚染の健康上の影響の長期調査から得られたサブセットです。このデータには、オハイオ州 スビューベンビル の 7 歳、8 歳、9 歳、10 歳の子供を対象に行った、喘鳴の状態の反復 2 値測定と、調査の初年に母親が喫煙していたかどうかの固定記録が含まれています。
- **workprog.sav**。体の不自由な人をより良い仕事に就かせようとする政府の事業プログラムに関する架空のデータ ファイルです。プログラムの参加者候補のサンプルが追跡されました。その中には、ランダムに選ばれてプログラムに登録された人と、そうでない人がいました。各ケースが別々のプログラム参加者を表します。
- **worldsales.sav** このデータ ファイルには、大陸および製品ごとの販売収益が含まれています。

# 注意事項

この情報は、世界各国で提供される製品およびサービス向けに作成されています。

IBMはこのドキュメントで説明する製品、サービス、機能は他の国では提供していない場合があります。現在お住まいの地域で利用可能な製品、サービス、および、情報については、お近くの IBM の担当者にお問い合わせください。IBM 製品、プログラム、またはサービスに対する参照は、IBM 製品、プログラム、またはサービスのみが使用することができることを説明したり意味するものではありません。IBM の知的所有権を侵害しない機能的に同等の製品、プログラム、またはサービスを代わりに使用することができます。ただし、IBM 以外の製品、プログラム、またはサービスの動作を評価および確認するのはユーザーの責任によるものです。

IBMは、本ドキュメントに記載されている内容に関し、特許または特許出願中の可能性があります。本ドキュメントの提供によって、これらの特許に関するいかなる権利も使用者に付与するものではありません。ライセンスのお問い合わせは、書面にて、下記住所に送ることができます。

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

2 バイト文字セット (DBCS) 情報についてのライセンスに関するお問い合わせは、お住まいの国の IBM Intellectual Property Department に連絡するか、書面にて下記宛先にお送りください。

神奈川県大和市下鶴間1623番14号 日本アイ・ビー・エム株式会社 法務・知的財産 知的財産権ライセンス渉外

**以下の条項は、イギリスまたはこのような条項が法律に反する他の国では適用されません。** International Business Machines は、明示的または黙示的に関わらず、第三者の権利の侵害しない、商品性または特定の目的に対する適合性の暗黙の保証を含むがこれに限定されない、いかなる保証なく、本出版物を「そのまま」提供します一部の州では、特定の取引の明示的または暗示的な保証の免責を許可していないため、この文が適用されない場合があります。

この情報には、技術的に不適切な記述や誤植を含む場合があります。情報については変更が定期的に行われます。これらの変更は本書の新版に追加されます。IBM は、本書に記載されている製品およびプログラムについて、事前の告知なくいつでも改善および変更を行う場合があります。

IBM 以外の Web サイトに対するこの情報内のすべての参照は、便宜上提供されているものであり、決してそれらの Web サイトを推奨するものではありません。これらの Web サイトの資料はこの IBM 製品の資料に含まれるものではなく、これらの Web サイトの使用はお客様の責任によるものとします。

IBM はお客様に対する一切の義務を負うことなく、自ら適切と考える方法で、情報を使用または配布することができるものとします。

本プログラムのライセンス取得者が (i) 別途作成されたプログラムと他のプログラム (本プログラムを含む) との間の情報交換および (ii) 交換された情報の相互利用を目的とした本プログラムに関する情報の所有を希望する場合、下記住所にお問い合わせください。

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

上記のような情報は、該当する条項および条件に従い、有料で利用できるものとします。

本ドキュメントに記載されている許可されたプログラムおよびそのプログラムに使用できるすべてのライセンス認証された資料は、IBM Customer Agreement、IBM International Program License Agreement、および当社とかわした同等の契約の条件に基づき、IBM によって提供されます。

IBM 以外の製品に関する情報は、それらの製品の供給業者、公開済みの発表、または公開で使用できるソースから取得しています。IBM は、それらの製品のテストは行っておらず、IBM 以外の製品に関連する性能、互換性、またはその他の要求については確認できません。IBM 以外の製品の性能に関する質問は、それらの製品の供給業者に通知する必要があります。

この情報には、日常の業務処理で用いられるデータや報告書の例が含まれています。できる限り詳細に説明するため、例には、個人、企業、ブランド、製品などの名前が使用されています。これらの名称はすべて架空のものであり、実際の企業で使用される名称および住所とは一切関係ありません。

この情報をソフトコピーでご覧になっている場合は、写真やカラーのイラストが表示されない場合があります。

## 商標

IBM、IBM ロゴ、および [ibm.com](http://www.ibm.com)、SPSS は、世界の多くの国で登録された IBM Corporation の商標です。IBM の商標の現在のリストは、<http://www.ibm.com/legal/copytrade.shtml> を参照してください。

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 のロゴ、Intel Inside、Intel Inside のロゴ、Intel Centrino、Intel Centrino のロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、および Pentium は、米国およびその他の国の Intel Corporation または関連会社の商標または登録商標です。

Java およびすべての Java ベースの商標およびロゴは、米国およびその他の国の Sun Microsystems, Inc. の商標です。

Linux は、米国およびその他の国における Linus Torvalds の登録商標です。

Microsoft、Windows、Windows NT、および Windows のロゴは、米国およびその他の国における Microsoft 社の商標です。

UNIX は、米国およびその他の国における The Open Group の登録商標です。

この製品は、WinWrap Basic (Copyright 1993-2007, Polar Engineering and Consulting, <http://www.winwrap.com>) を使用します。

その他の製品名およびサービス名等は、IBM または他の会社の商標です。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。



# 索引

- 完全条件指定
  - [多重代入] での, 21
- 欠損指示変数
  - 欠損値分析, 7
- 度数分布表
  - 欠損値分析, 7
- 欠損値分析, 2
  - 記述統計量, 7
  - 回帰, 11
  - 方法, 8
  - EM(E), 10
  - MCAR 検定, 9
  - コマンドの追加機能, 13
  - 期待値の最大化, 12
  - 欠損値の代入, 8
  - 統計量の推定, 8
  - パターン, 5
- 単調代入
  - [多重代入] での, 21
- 多重代入, 15, 27, 31
  - 欠損データ値の代入, 18
  - パターン分析, 17
- 指示変数
  - 欠損値分析, 7
- 標準偏差
  - 欠損値分析, 7
- 正規変数
  - 欠損値分析, 11
- 相関係数
  - 欠損値分析, 10-11
- 不一致
  - 欠損値分析, 7
- 共分散
  - 欠損値分析, 10-11
- 欠損値
  - 1変量統計量, 7
- 商標, 104
- 回帰
  - 欠損値分析, 11
- 平均
  - 欠損値分析, 7, 10-11
- 残差
  - 欠損値分析, 11
- 1変量統計量
  - 欠損値分析, 44
- EM(E)
  - 欠損値分析, 10
- FCS 収束グラフ
  - 多重代入による, 80
- Little の MCAR 検定, 9
  - 欠損値分析, 2
- Little の MCAR 検定
  - 欠損値分析, 53
- MCAR 検定
  - 欠損値分析, 2
- MCAR 検定
  - 欠損値分析, 53
- t 検定
  - 欠損値分析, 7
- オプション
  - 多重代入, 38
- カテゴリの集計
  - 欠損値分析, 7
- カテゴリの集計
  - 欠損値分析, 46
- 欠損値
  - 1変量統計量, 44
  - 欠損値パターン, 52
  - 欠損値分析, 41
    - 記述統計量, 41
    - パターン, 50
  - ケースの集計
    - 欠損値分析, 5
  - ケースの並べ替え
    - 欠損値分析, 5
- サンプル ファイル
  - 位置, 92
- スチューデント t 検定
  - 欠損値分析, 45
- スチューデント t 検定
  - 欠損値分析, 11
- 多重代入, 55
  - FCS 収束グラフ, 80
  - オプション, 38
  - 記述統計量, 65, 74
  - 欠損値の全体の要約, 56
  - 欠損値パターン, 58
  - 制約, 74
  - 代入結果, 63
  - 代入指定, 62
  - プールされた結果, 80
  - プールされた推定値, 86
  - 変数の要約, 57

モデル, 64

不完全データ

「欠損値分析」を参照, 2

欠損データ値を代入, 18

代入方法, 21

出力, 26

制約, 23

法律に関する注意事項, 103

t 検定

欠損値分析, 45

反復の記述

[多重代入]での, 26

極値の度数

欠損値分析, 7

パターン分析, 17

プールされた結果

多重代入による, 80

プールされた推定値

多重代入による, 86

ペアごとの削除

欠損値分析, 2

リストごとの削除

欠損値分析, 2