トレーニング コース

IBM SPSS Statistics データ加工

演習解答例

IBM SPSS Statistics V20
2012/09

この資料は研修用教材として作成したものです。

この資料は2012年9月1日現在のものであり、将来この資料の全体または一部につき変更する場合があります。

本書の内容についてお気づきの点がございましたら、下記までお知らせください。

〒103-8510　東京都中央区日本橋箱崎町19-21

TEL. 0120-105553

日本アイ・ビー・エム株式会社

演習解答例

第2章　数値データの計算

ここでは、ある金融機関の従業員のデータ（BANK.sav）を使います。このデータには、従業員コード（id）、性別（gender）、初任給（salbeg）、現在の給与（salnow）、などの情報が記録されています。

数値データの変換

1. IBM SPSS Statisticsを起動してBANK.savファイルを開いてください。

☞ WindowsのスタートメニューからIBM SPSS Statisticsを起動します。
ファイルメニューの**開く**の**データ**を選択します。
C:¥train¥DataMani¥BANK.savを選択し、**開く**ボタンをクリックします。

2. predictという名前の新しい変数を作成してください。この変数は、回答者の年齢（age）と教育年数（edlevel）に基づく現在の収入の予測値です。次の式を使ってください。

1592 * edlevel + 25 * age – 8600

☞ **変換**メニューの**変数の計算**を選択します。
目標変数ボックスに**predict**、**数式**に**1529 * edlevel + 25 * age – 8600**と入力します。
OKボタンをクリックします。

図E.1　　新しい変数を計算するダイアログボックス

3. 初任給（salbeg）と現在の給与（salnow）の算術平均をあらわす変数を作成してください。

☞ **変換**メニューの**変数の計算**を選択します。
目標変数ボックスに**avesal**（任意の変数名）、**数式**に**(salbeg + salnow) / 2**と入力します。
OKボタンをクリックします。

4. MEAN関数を使って、3.と同様の平均をあらわす変数を作成してください。２つの結果に違いはありますか？

☞ **変換**メニューの**変数の計算**を選択します。
目標変数ボックスに**meansal**（任意の変数名）、数式に**MEAN(salbeg,salnow)**を入力します。
OKボタンをクリックします。

☞ **分析**メニューの**記述統計**の**記述統計**を選択します。
avesalと**meansal**を**変数**リストボックスに移動し、**OK**ボタンをクリックします。

記述統計量

	度数	最小値	最大値	平均値	標準偏差
avesal	474	5190.00	42996.00	10287.1308	4858.37139
meansal	474	5190.00	42996.00	10287.1308	4858.37139
有効なｹｰｽの数 (ﾘｽﾄごと)	474				

図E.2　　２つの平均を比較する

統計量の値がすべて等しいので、２つの変数には違いがないと考えられます。

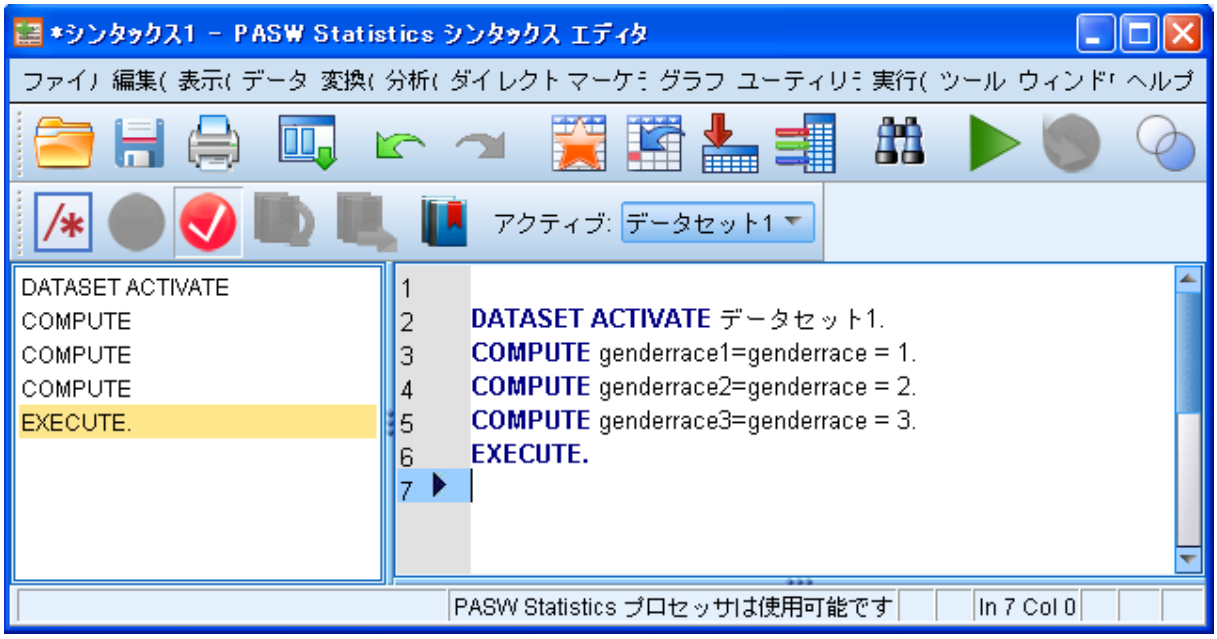
5. 性別と人種をあらわすgenderraceという質的データから、ダミー変数を作成してください。コード化の例は以下を参考にしてください。

genderrace	genderrace1	genderrace2	genderrace3
1（白人男性）	1	0	0
2（白人以外男性）	0	1	0
3（白人女性）	0	0	1
4（白人以外女性）	0	0	0

☞ **変換**メニューの**変数の計算**を選択します。
目標変数ボックスに**genderrace1**、**数式**ボックスに**genderrace = 1**と入力します。
OKボタンをクリックします。

図E.3　　genderrace1を作成

☞ 同じ要領で、**genderrace2（数式：genderrace = 2）**、**genderrace3（数式：genderrace = 3）**
を作成します。シンタックスの貼り付けを使って作成してもよいでしょう。



図E.4　　シンタックスを使ったダミー変数の作成

図E.5　　作成したダミー変数を表示

第3章 日付データと文字データの操作

文字型変数の変換

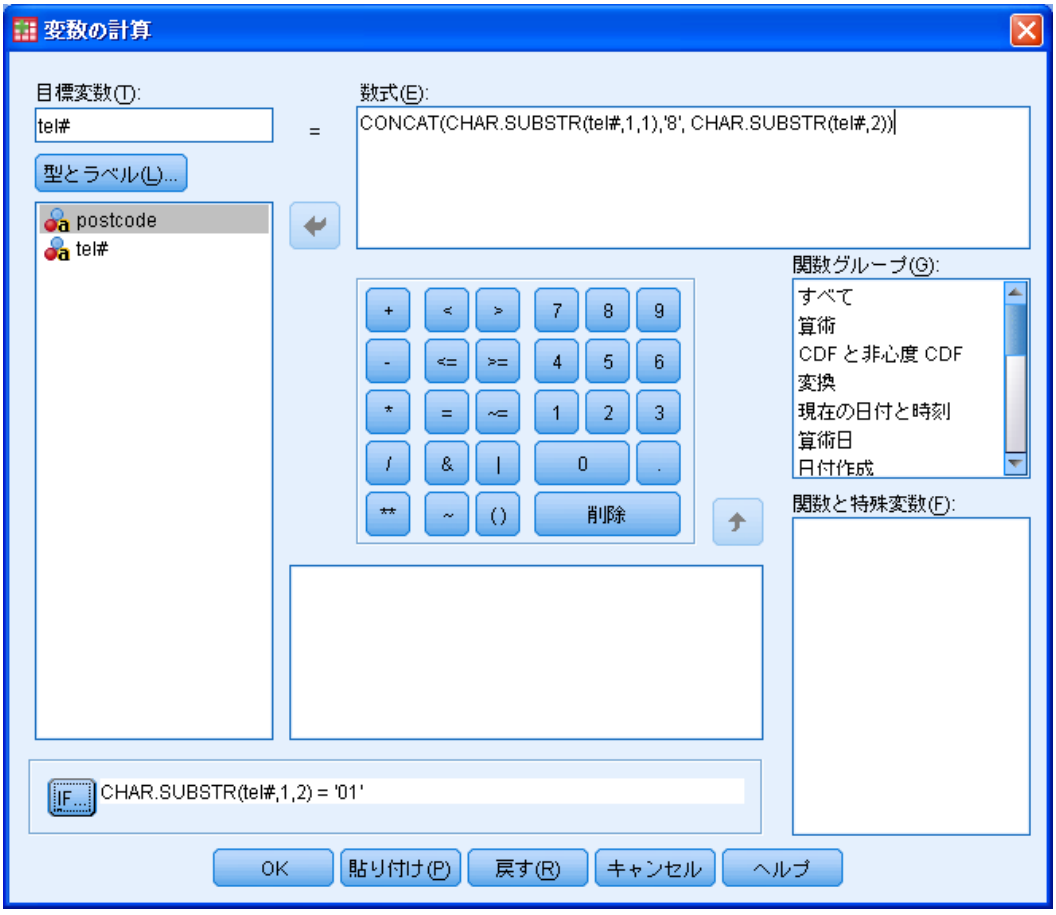
1. POST&TEL.savファイルを開いてください。この郵便番号コードを示すpostcodeの値は正しいものではありません。最初の文字**D**の後に**E**を追加しなければなりません。変数の計算手続きと文字型関数を組み合わせて、イギリスの正しい郵便番号コードをあらわす新しい変数newpostを作成してください。

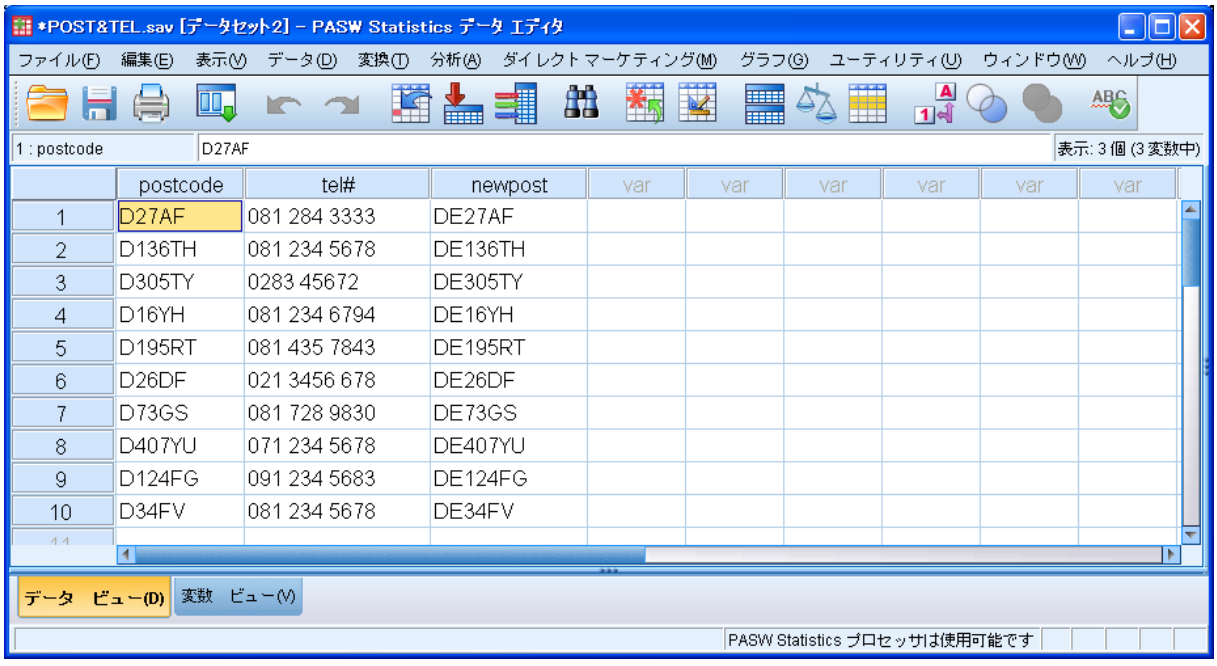
☞ **ファイルを開く**ボタンをクリックし、**POST&TEL.sav**を選択し、**開く**ボタンをクリックします。
変換メニューの**変数の計算**をクリックします
目標変数ボックスに**newpost**と入力します。
型とラベルボタンをクリックして型の**文字型**を選択し、**続行**ボタンをクリックします。
文字式ボックスに
CONCAT(CHAR.SUBSTR(postcode,1,1),'E',CHAR.SUBSTR(postcode,2))と入力します。
OKボタンをクリックします。

図E.6　　newpostを作成する

2. tel#として保存されている電話番号のうち、01で始まるロンドンの市外局番の値は正しいものではありません。変数の計算手続きで、条件を指定し、文字型関数を使ってロンドンの電話番号の最初の**0**の直後に**8**を挿入してください（01を081へ変更）。

☞ **変換**メニューの**変数の計算**を選択します。
目標変数ボックスに**tel#**と入力します。
文字式ボックスに、**CONCAT(CHAR.SUBSTR(tel#,1,1),'8',**
CHAR.SUBSTR(tel#,2))と入力します。
IFボタンをクリックし、**IF条件を満たしたケースを含む**を選択します。
条件式**CHAR.SUBSTR(tel#,1,2) = '01'**と入力し、**続行**ボタンをクリックします。
OKボタンをクリックします。





	postcode	tel#	newpost	var	var	var	var	var	var
1	D27AF	081 284 3333	DE27AF						
2	D136TH	081 234 5678	DE136TH						
3	D305TY	0283 45672	DE305TY						
4	D16YH	081 234 6794	DE16YH						
5	D195RT	081 435 7843	DE195RT						
6	D26DF	021 3456 678	DE26DF						
7	D73GS	081 728 9830	DE73GS						
8	D407YU	071 234 5678	DE407YU						
9	D124FG	091 234 5683	DE124FG						
10	D34FV	081 234 5678	DE34FV						

図E.7　　計算ダイアログと実行結果

日付型データの変換

3. HOSPITAL.savファイルを開いてください。このファイルには４人の患者の情報が記録されています。

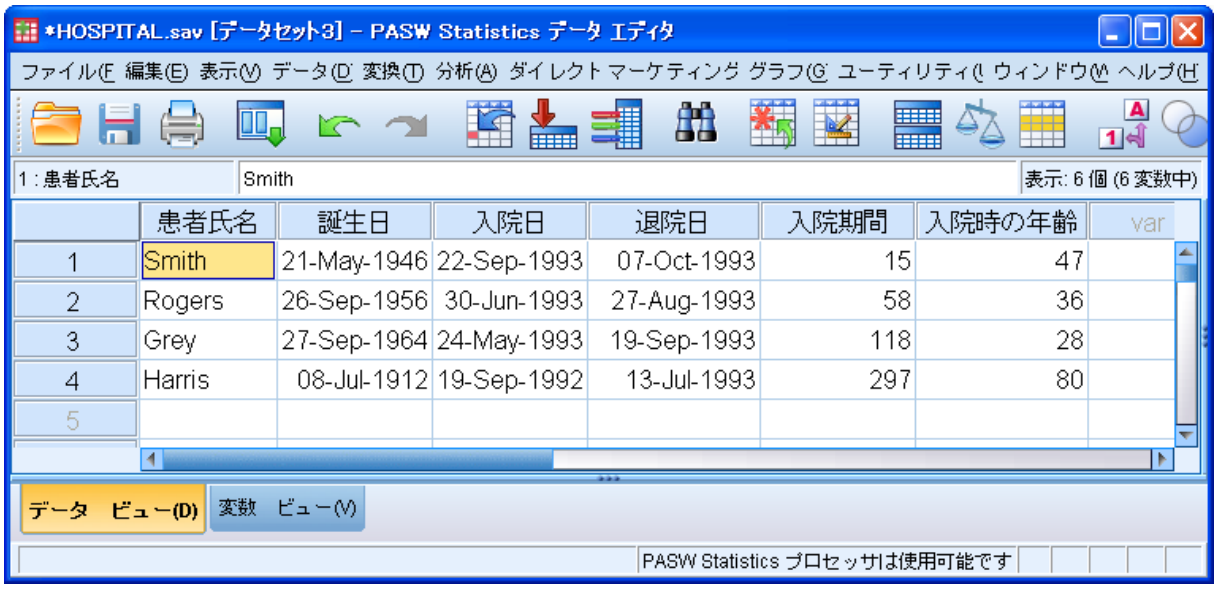
☞ **ファイル**メニューの**開く**の**データ**を選択します。
HOSPITAL.savを選択し、**開く**ボタンをクリックします。

4. 入院期間（日数単位）を計算してください。

☞ **変換**メニューの**日付と時刻ウィザード**を選択します。
日付と時刻ウィザードで**日付と時刻で計算**を選択し、**次へ**ボタンをクリックします。
日付と時刻ウィザード（ステップ１／３）で、**２つの日付間の時間単位数の計算**を選択し、**次へ**ボタンをクリックします。
日付と時刻ウィザード（ステップ２／３）で、**日付１**に**退院日**、**引く日付２**に**入院日**を移動し、**単位**を**日**に変更します。
次へボタンをクリックします。
日付と時刻ウィザード（ステップ３／３）で、**変数**ボックスに**入院期間**と入力し、**完了**ボタンをクリックします。

5. 各患者の入院時の年齢（満年齢）を計算してください。

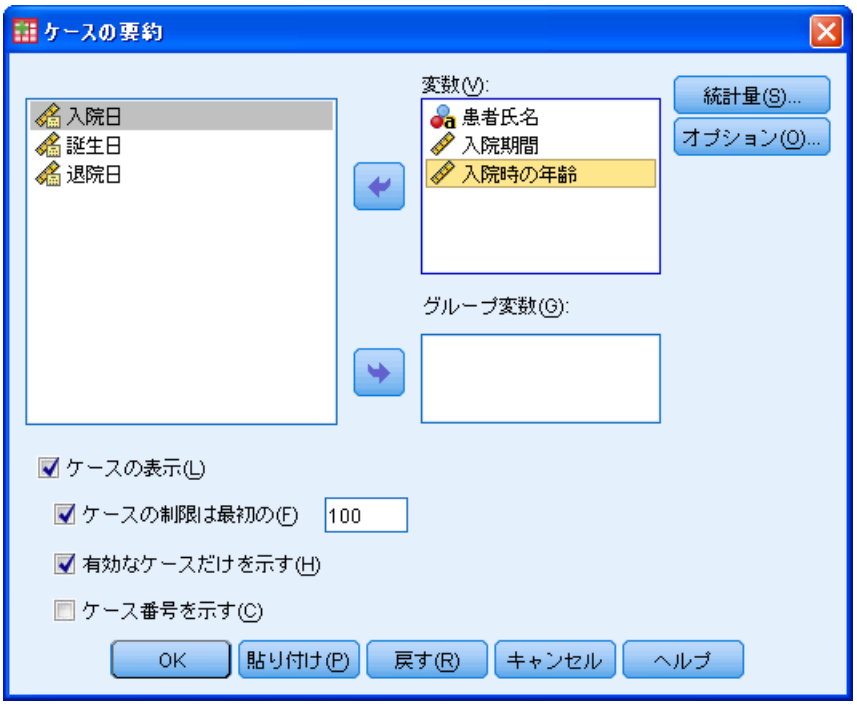
☞ **変換**メニューの**日付と時刻ウィザード**を選択します。
日付と時刻ウィザードで**日付と時刻で計算**を選択し、**次へ**ボタンをクリックします。
日付と時刻ウィザード（ステップ１／３）で、**２つの日付間の時間単位数の計算**を選択し、**次へ**ボタンをクリックします。
日付と時刻ウィザード（ステップ２／３）で、**日付１**に**入院日**、**引く日付２**に**誕生日日**を移動します（**単位**はデフォルトの**年**を使用します）。
次へボタンをクリックします。
日付と時刻ウィザード（ステップ３／３）で、**変数**ボックスに**入院時の年齢**と入力し、**完了**ボタンをクリックします。



図E.8　　入院時の年齢

6.　患者の氏名、入院日数、入院時の年齢を記述する出力を作成してください。

☞　**分析**メニューの**報告書**の**ケースの要約**を選択します。
患者氏名、入院期間、入院時の年齢を**変数**リストボックスに移動します。
OKボタンをクリックします。



ケースの集計[a]

		患者氏名	入院期間	入院時の年齢
1		Smith	15	47
2		Rogers	58	36
3		Grey	118	28
4		Harris	297	80
合計	度数	4	4	4

a. 最初の 100 のｹｰｽに制限されています。

図E.9　　ケースの集計手続きと結果

第4章　ケースの選択とファイルの分割

CUSTOMER.savは金融機関の顧客のデータです。このデータには、顧客番号（id）、収入（income）、性別（gender）、地域（region）、婚姻状況（marital）、子供の数（child）、車の所有（car）、担保の有無（mortgage）、口座の種類（acct）が記録されており、358のケースがあります。

ケースの選択

1. CUSTOMER.savを開いてください。

☞ **ファイル**メニューの**開く**の**データ**を選択します。
CUSTOMER.savを選択し、**開く**ボタンをクリックします。

2. 収入（income）の平均値を求めてください。

☞ **分析**メニューの**記述統計**の**記述統計**を選択します。
incomeを**変数**リストボックスに移動し、**OK**ボタンをクリックします。

記述統計量

	度数	最小値	最大値	平均値	標準偏差
収入	358	5960	61476	27124.01	12522.262
有効なｹｰｽの数 (ﾘｽﾄごと)	358				

図E.10　記述統計量

3. ケースの選択ダイアログボックスを使用して、収入が平均値（2.で求めた値）以上のケースを選択してください。

☞ **データ**メニューの**ケースの選択**を選択します。
IF条件が満たされるケースを選択して**IF**ボタンをクリックします。
条件式**income >= 27124.01**を入力して**続行**ボタンをクリックします。
OKボタンをクリックします。

図E.11　ケースの選択

4. 選択されたケースには、男性と女性のどちらが多いでしょうか？婚姻状況ではどのカテゴリが多いでしょうか？

☞ **分析**メニューの**記述統計**の**度数分布表**を選択します。
gender、**marital**を**変数**リストボックスに移動し、**OK**ボタンをクリックします。

性別

		度数	パーセント	有効パーセント	累積パーセント
有効	女性	80	54.1	54.1	54.1
	男性	68	45.9	45.9	100.0
	合計	148	100.0	100.0	

婚姻状況

		度数	パーセント	有効パーセント	累積パーセント
有効	既婚	85	57.4	57.4	57.4
	未婚	63	42.6	42.6	100.0
	合計	148	100.0	100.0	

図E.12　度数分布表手続きと実行結果

男性より女性の方が多いようです。また、未婚者より既婚者のほうが多いようです。

5. ケースの選択を解除してください。

☞ **データ**メニューの**ケースの選択**を選択します。
すべてのケースを選択して**OK**ボタンをクリックします。

ファイルの分割

6. ファイルの分割ダイアログボックスを使用して、地域（region）でグループを作成し比較するように指定をしてください。

☞ **データ**メニューの**ファイルの分割**を選択します。
グループの比較を選択します。
regionを**グループ化変数**リストボックスに移動します。
OKボタンをクリックします。

図E.13　ファイルの分割ダイアログボックス

7. 地域による収入の違いを比較できるような手続きを実行してください。

☞ **分析**メニューの**記述統計**の**記述統計**を選択します。
incomeを**変数**リストボックスに移動し、**OK**ボタンをクリックします。

記述統計量

地域		度数	最小値	最大値	平均値	標準偏差
郊外	収入	63	9785	51468	26211.25	12366.113
	有効なケースの数 (リストごと)	63				
市街地	収入	110	5960	61476	24762.93	10575.899
	有効なケースの数 (リストごと)	110				
都市	収入	145	6531	60635	28343.70	13218.709
	有効なケースの数 (リストごと)	145				
農村	収入	40	8005	60176	30633.20	14116.449
	有効なケースの数 (リストごと)	40				

図E.14 実行結果

収入の平均値を比較すると、収入が最も多いのは農村部、最も少ないのは市街地であることがわかります。本当に違いがあるのかを調べるためには、さらに分析が必要です。

第5章 グループ集計とケースの重み付け

CUSTOMER.savのデータには、同じid（顧客番号）を持つケースが複数入力されています。このデータをグループ集計し、1ケースが1個人を表すように集計の単位を指定します。

1. データのグループ集計機能を使用して、id（顧客番号）ごとに次の集計を行ってケースを作成します。

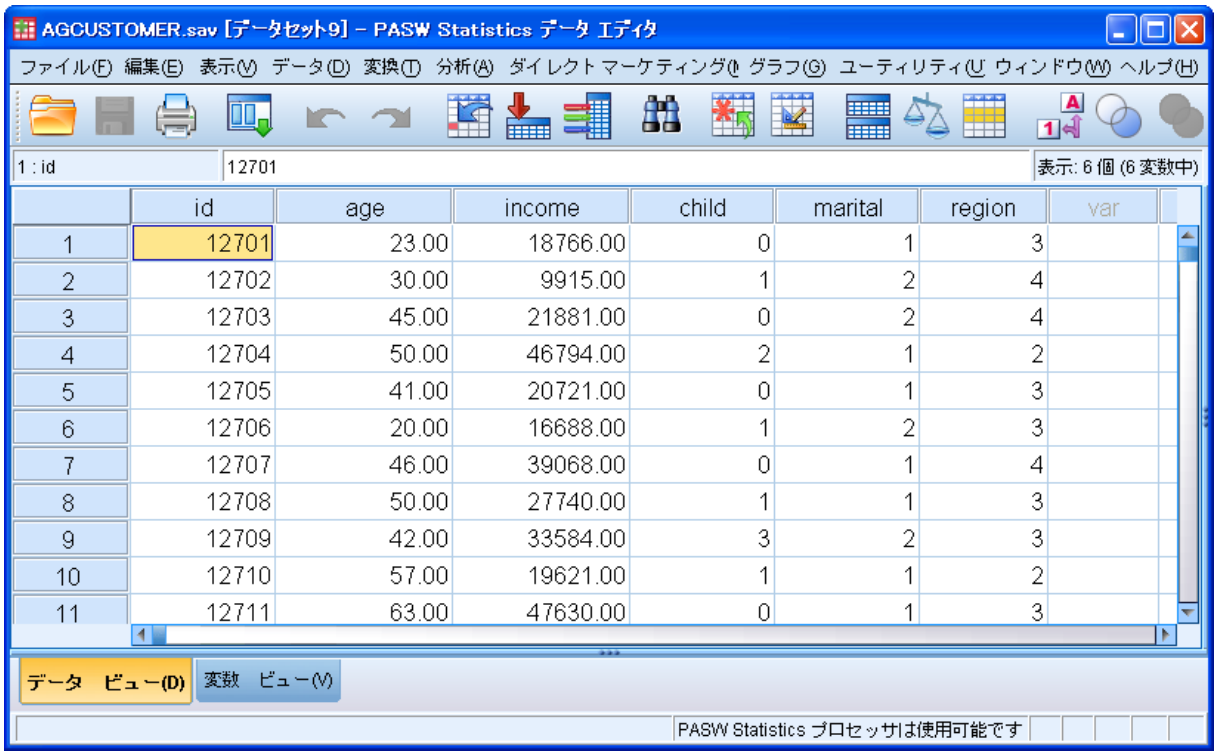
変数名	集計方法
age	age（年齢）の平均値
income	income（収入）の平均値
child	child（子供の数）の最大値
marital	marital（婚姻状況）の最後の値
region	region（地域）の最後の値

☞ **ファイル**メニューの**開く**の**データ**を選択します。
CUSTOMER.savを選択し、**開く**ボタンをクリックします。
データメニューの**グループ集計**を選択します。
idを**ブレーク変数**リストボックスに、**age**、**income**、**child**、**marital**、**region**を**変数の集計**リストボックスに移動します。
関数ボタンを使用して、集計方法を上記の指定内容に変更します。
変数名とラベルボタンを使用して、変数名を上記名に変更します。

2. 集計ファイルをAGCUSTOMER.savとして保存します。保存先はC:\train\DataManiフォルダを指定します。

☞ **集計変数のみを含む新しいデータファイルを作成する**を選択します。
ファイルボタンをクリックし、**C:\train\DataMani**フォルダで、**ファイル名**テキストボックスに**AGCUSTOMER.sav**と入力し、**保存**ボタンをクリックします。
OKボタンをクリックします。
データのグループ集計：一致する名前の警告ダイアログボックスで、**上書き**ボタンをクリックします。


データのグループ集計
acct
age
car
child
filter_$
gender
income
marital
mortgage
region
ブレーク変数(B):
id
集計変数
変数の集計(S):
age = MEAN(age)
income = MEAN(income)
child = MAX(child)
marital = LAST(marital)
region = LAST(region)
関数(F)...
変数名とラベル(N)...
ケースの数(C)
名前(M):
N_BREAK
保存
アクティブなデータセットに集計変数を追加(D)
集計変数のみを含む新しいデータセットを作成する(E)
データセット名(D):
集計変数のみを含む新しいデータ ファイルを作成する(W)
ファイル(L)...
C:\train\DataMani\AGCUSTOMER.sav
非常に大きなデータセットのオプション
ファイルは既にブレーク変数で並べ替えられている(A)
グループ集計前にファイルを並べ替え(G)
OK
貼り付け(P)
戻す(R)
キャンセル
ヘルプ



	id	age	income	child	marital	region	var
1	12701	23.00	18766.00	0	1	3	
2	12702	30.00	9915.00	1	2	4	
3	12703	45.00	21881.00	0	2	4	
4	12704	50.00	46794.00	2	1	2	
5	12705	41.00	20721.00	0	1	3	
6	12706	20.00	16688.00	1	2	3	
7	12707	46.00	39068.00	0	1	4	
8	12708	50.00	27740.00	1	1	3	
9	12709	42.00	33584.00	3	2	3	
10	12710	57.00	19621.00	1	1	2	
11	12711	63.00	47630.00	0	1	3	

図E.15　データをidで集計

3. 次に、CUSTOMER.savを、region（居住地域）の単位で集計します。次の集計をダイアログボックスで指定してください。

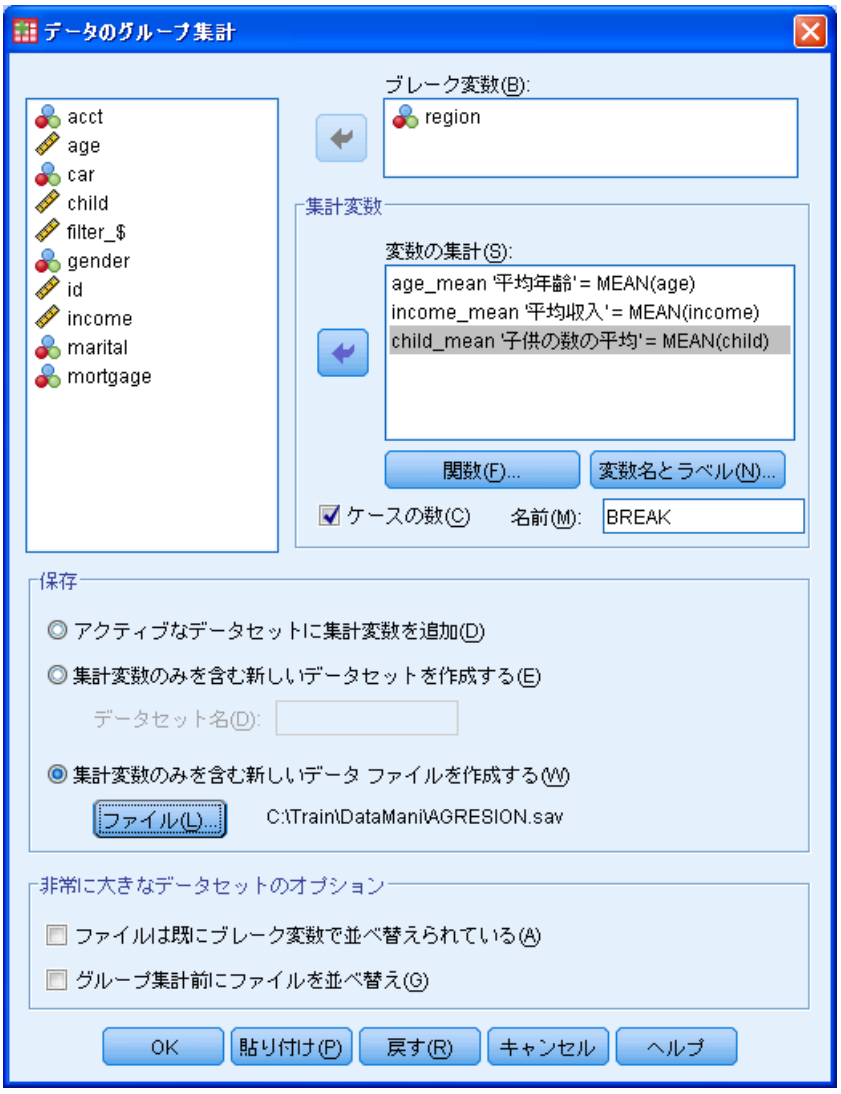
変数名（変数ラベル）	**集計方法**
age_mean（平均年齢）	最適な方法を選択
income_mean（平均収入）	最適な方法を選択
child_mean（子供の数の平均）	最適な方法を選択
break（地域の人数）	最適な方法を選択

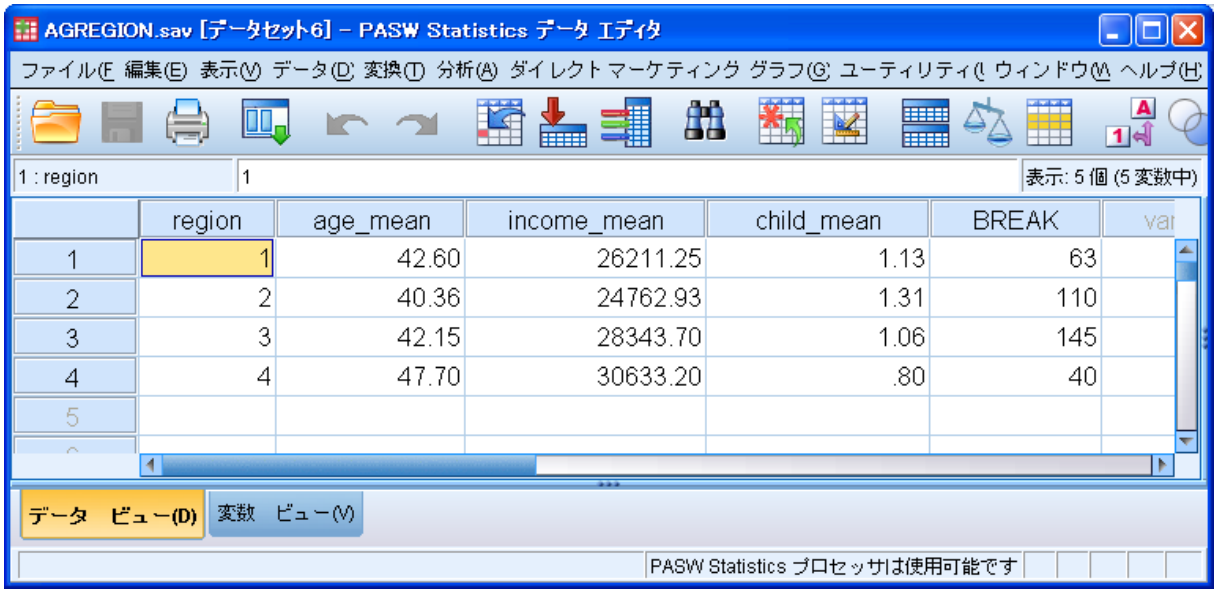
☞ **データ**メニューの**グループ集計**を選択します（必要に応じて**戻す**ボタンをクリックします）。
regionを**グループ変数**リストボックスに、**age**、**income**、**child**を**変数の集計**リストボックスに移動します。
変数名とラベルボタンをクリックして上記の変数ラベルを入力します。
ケースの数を選択して**名前**ボックスに**BREAK**と入力します。

4. ファイル名をAGRESION.savとして保存します。保存先は、C:\train\DataManiフォルダを指定します。

☞ **集計変数のみを含む新しいデータファイルを作成する**を選択します。**ファイル**ボタンをクリックし、**C:\train\DataMani**フォルダで、**ファイル名**テキストボックスに**AGRESION.sav**と入力し、**保存**ボタンをクリックします。

OKボタンをクリックします。





図E.16　データのグループ集計ダイアログと集計結果

第6章 ファイルの結合 ― ケースの追加

銀行の顧客に関する97年度（97DATA.sav）と98年度（98DATA.sav）のデータがあります。データに入力されている情報は、顧客番号（id）、口座の種類（acct）、口座開設時の残高（openbal）、現在の残高（currbal）です。この２つのデータの結合を行い、１つのデータファイルを作成します。さらに、そのデータを顧客番号で集計したファイルを作成します。

1. 97年度のデータを開きます。ファイルの結合ダイアログボックスを使用して、98年度のデータファイルを追加します。ケースソースを示すための新しい変数sourceを作成します。

☞ **ファイル**メニューの**開く**の**データ**を選択します。
97DATA.savを選択し、**開く**ボタンをクリックします。
データメニューの**ファイルの結合**の**ケースの追加**を選択します。
外部SPSSデータファイルを選択し、**参照**ボタンをクリックします。
98DATA.savを選択し、**開く**ボタンをクリックします。
続行ボタンをクリックします。
変数としてソースケースを示すを選択し、変数名を**SOURCE**に変更します。
OKボタンをクリックします。

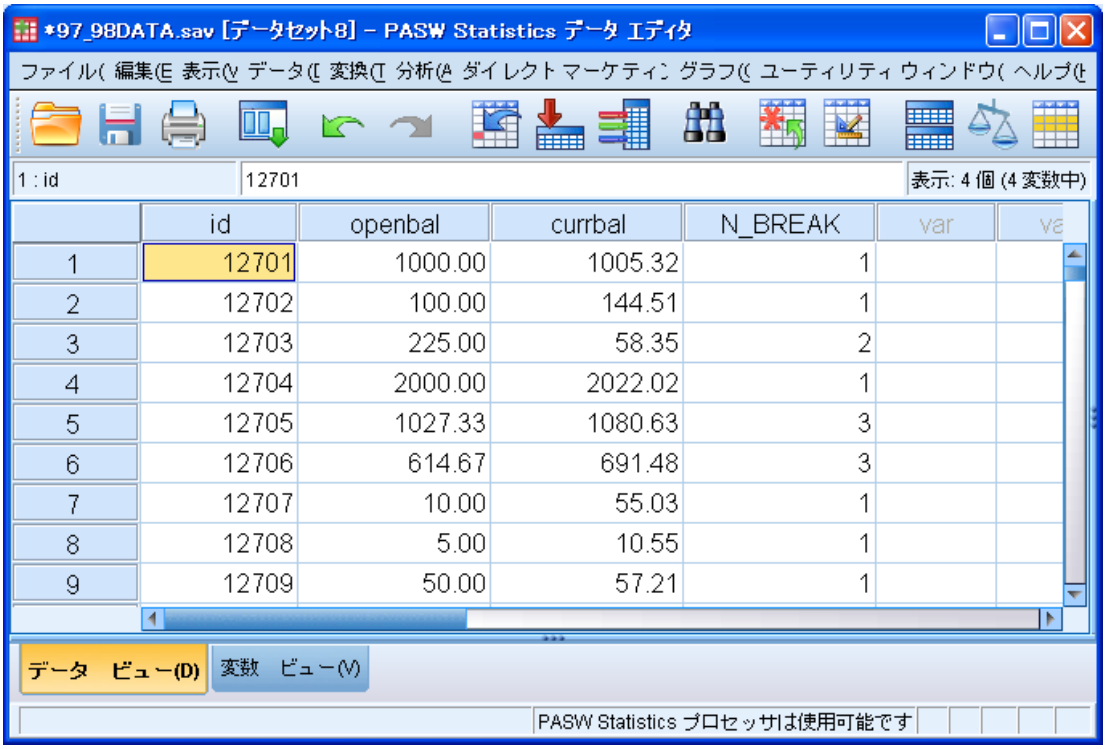
2. グループ集計ダイアログボックスを使って集計ファイルを作成します。
ブレーク変数には、顧客番号（id）を投入し、ブレークグループのケース数をN_BREAKの変数名で保存します。

変数名	**集計方法（変数ラベル）**
openbal	口座開設時残高の平均
currbal	現在の残高の平均

☞ **データ**メニューの**グループ集計**を選択します。
idを**グループ変数**リストボックスに、**openbal**、**currbal**を**変数の集計**リストボックスに移動します。
変数名とラベルボタンをクリックして上記の変数名､変数ラベルを入力します。
ケースの数を選択します。

3. 新しいデータファイルを作成し、保存します。C:\train\DataManiフォルダに、97_98DATA.savのファイル名で保存します。

☞ **集計変数のみを含む新しいデータファイルを作成する**を選択します。
ファイルボタンをクリックし、**C:\train\DataMani**フォルダで、**ファイル名**テキストボックスに**97_98DATA.sav**と入力し、**保存**ボタンをクリックします。
OKボタンをクリックします。



図E.17　97_98Data.savファイル

第7章 ファイルの結合 ― 変数の追加

１対１の結合

第5章の演習問題で作成したAGCUSTOMER.savと第6章の演習問題で作成した97_98DATA.savの２つのデータの結合を行います。

1. AGCUSTOMER.sav ファイルと97_98DATA.savファイルを検討し、キー変数を決定します。

☞ **ファイル**メニューの**開く**の**データ**を選択します。
AGCUSTOMER.savを選択し、**開く**ボタンをクリックします。
データビュー、**変数ビュー**を表示して内容を確認します。
同じように、**97_98DATA.sav**も確認します。

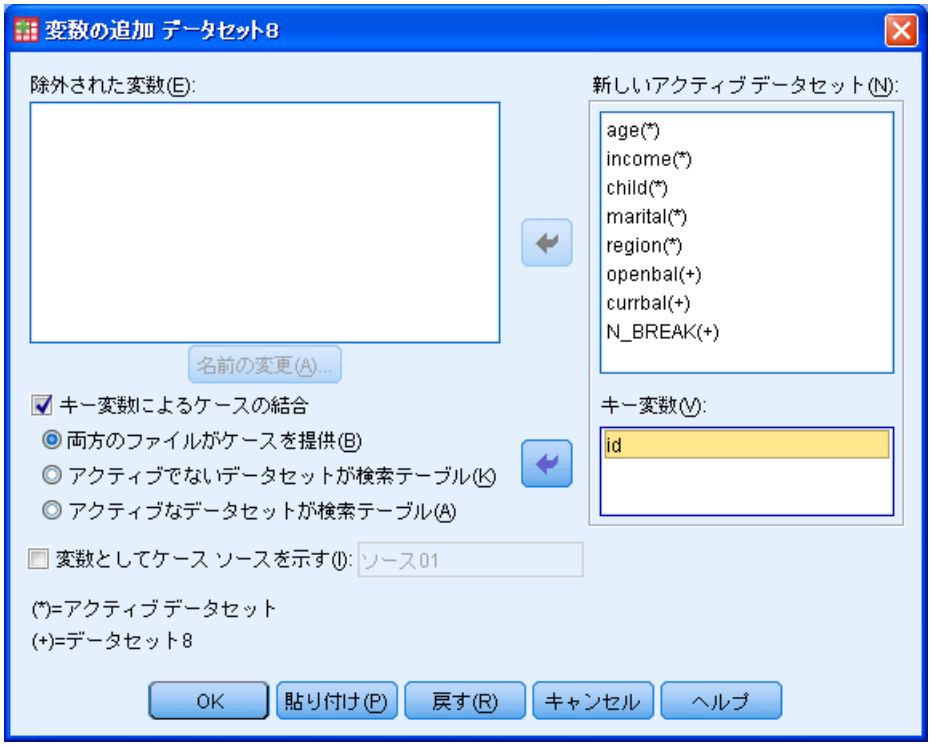
どちらのデータもid（顧客番号）ごとにケースが集計されています。これをキー変数として使うことにします。

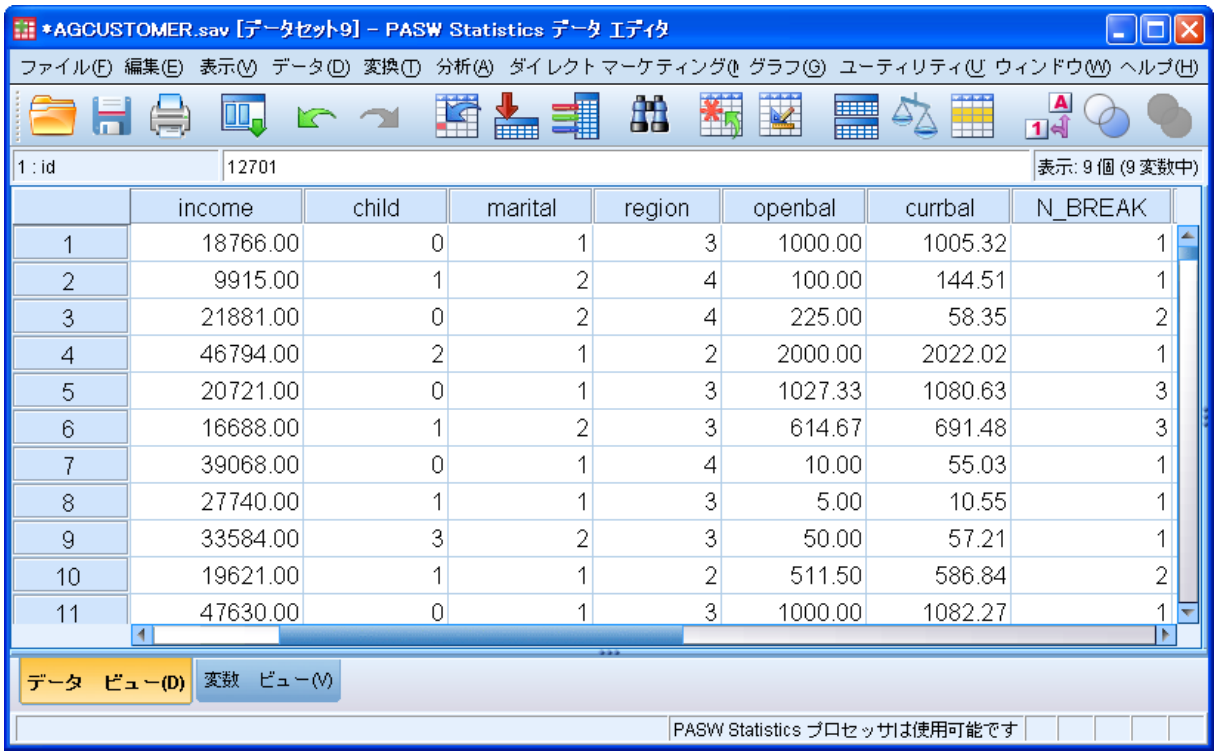
2. ２つのファイルをキー変数でソートします。

☞ **AGCUSTOMER.sav**が表示されているデータエディタで、**データ**メニューの**ケースの並び替え**を選択します。
idを**並べ替え**ボックスに移動し、**OK**ボタンをクリックします。
同じように、**97_98DATA.sav**も**id**で並び替えます。

3. 変数の追加ダイアログボックスを使用して、AGCUSTOMER.savと97_98DATA.savを結合します。

☞ **AGCUSTOMER.sav**が表示されているデータエディタで、**データ**メニューの**ファイルの結合**の**変数の追加**を選択します。
開いているデータセットを選択し、**97_98DATA.sav**を選択して**続行**ボタンをクリックします。
キー変数によるケースの結合を選択します。
除外された変数リストボックスの**id**を**キー変数**リストボックスに移動します。
OKボタンをクリックします。





	income	child	marital	region	openbal	currbal	N_BREAK
1	18766.00	0	1	3	1000.00	1005.32	1
2	9915.00	1	2	4	100.00	144.51	1
3	21881.00	0	2	4	225.00	58.35	2
4	46794.00	2	1	2	2000.00	2022.02	1
5	20721.00	0	1	3	1027.33	1080.63	3
6	16688.00	1	2	3	614.67	691.48	3
7	39068.00	0	1	4	10.00	55.03	1
8	27740.00	1	1	3	5.00	10.55	1
9	33584.00	3	2	3	50.00	57.21	1
10	19621.00	1	1	2	511.50	586.84	2
11	47630.00	0	1	3	1000.00	1082.27	1

図E.18　変数の追加ダイアログと実行結果

4.　結合後のファイルを、FULLDATA.savのファイル名で保存します。

☞　**ファイル**メニューの**名前を付けて保存**を選択します。
ファイル名に**FULLDATA.sav**と入力して**保存**ボタンをクリックします。

検索テーブルによる結合

1対1の結合で作成したFULLDATA.savに、検索テーブルによる結合を使用して、第5章の演習問題で作成したAGRESION.savを結合します。結合したデータで、各顧客のデータと地域ごとの平均値とを比較します。

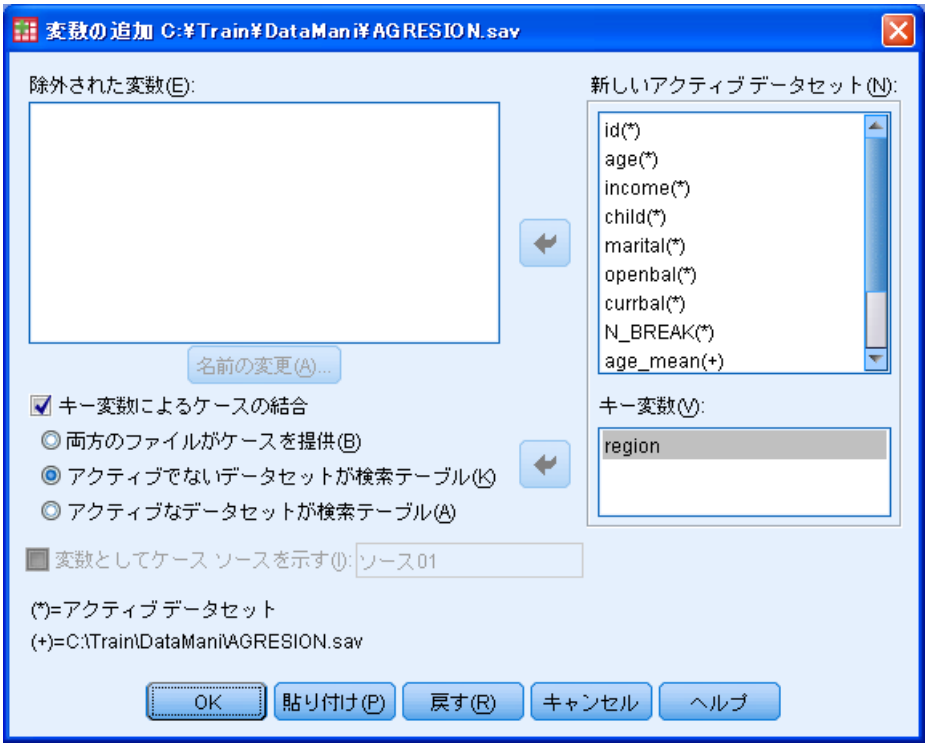
5. １対１の結合で作成したFULLDATA.savを開きます。キー変数を決定し、ソートしておきます。

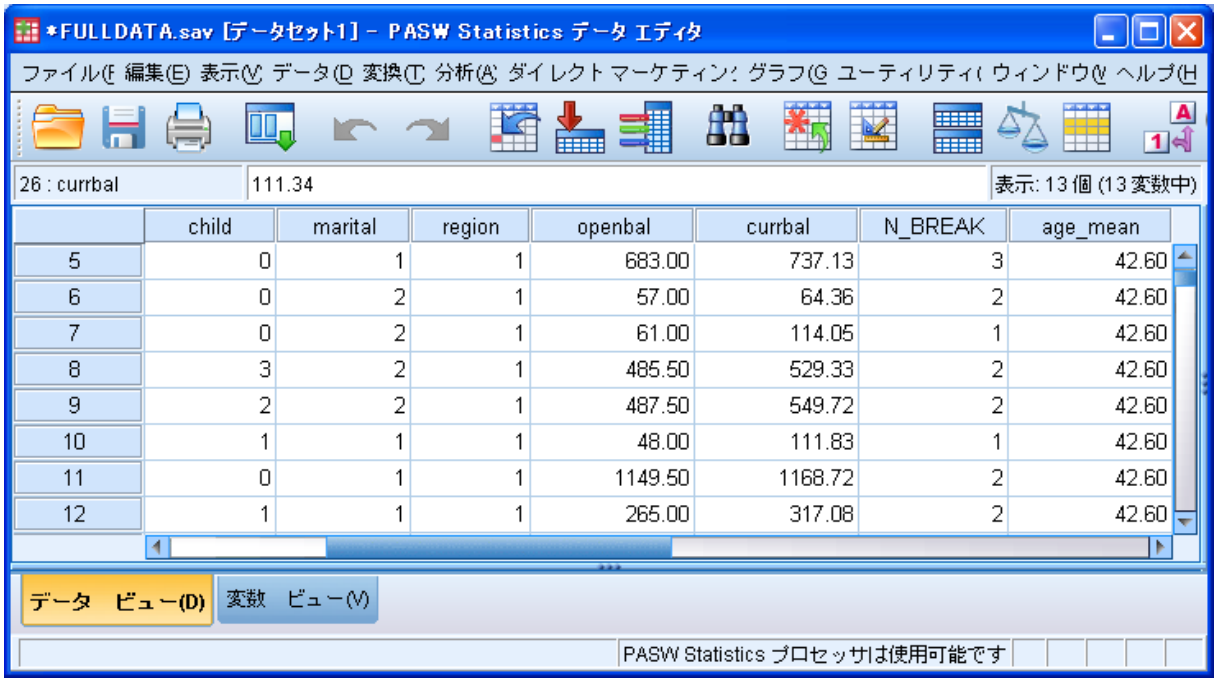
AGRESION.savはregion（居住地域）で集計されています。FULLDATA.savにもregionがあります。regionをキー変数として、AGRESION.savを検索テーブルとしたファイルの結合を行います。

☞ **ファイル**メニューの**開く**の**データ**を選択します。
FILLDATA.savを選択し、**開く**ボタンをクリックします。
データメニューの**ケースの並び替え**を選択します。
regionを**並び替え**リストボックスに移動し、**OK**ボタンをクリックします。

6. 変数の追加ダイアログボックスを開きます。結合するファイル（AGRESION.sav）を選択します。外部ファイルが検索テーブルであることを指定します。

☞ **FULLDATA.sav**が表示されているデータエディタで、**データ**メニューの**ファイルの結合**の**変数の追加**を選択します。
外部SPSSデータファイルを選択し、**参照**ボタンをクリックして
AGRESION.savを選択し、**開く**ボタンをクリックします。
続行ボタンをクリックします。
キー変数によるケースの結合を選択し、**アクティブでないデータセットが検索テーブル**を選択します。
除外された変数リストボックスの**region**を**キー変数**リストボックスに移動します。
OKボタンをクリックします。





	child	marital	region	openbal	currbal	N_BREAK	age_mean
5	0	1	1	683.00	737.13	3	42.60
6	0	2	1	57.00	64.36	2	42.60
7	0	2	1	61.00	114.05	1	42.60
8	3	2	1	485.50	529.33	2	42.60
9	2	2	1	487.50	549.72	2	42.60
10	1	1	1	48.00	111.83	1	42.60
11	0	1	1	1149.50	1168.72	2	42.60
12	1	1	1	265.00	317.08	2	42.60

図E.19　変数の追加ダイアログと実行結果

7. 結合したファイルを使用して、各顧客の収入と地域ごとの収入の平均値とを比較します。

収入が地域の平均より低いことをあらわす変数を作成し、その度数分布表を作成して、収入が地域の平均より低い人がどれくらいいるのか調べます。

☞ **変換**メニューの**変数の計算**を選択します。
目標変数ボックスに**low_income**（任意の名前）、**数式**に**income < income_mean**と入力し、**OK**ボタンをクリックします。

図E.20　変数の計算ダイアログボックス

☞ **分析**メニューの**記述統計**の**度数分布表**を選択します。
low_incomeを**変数**リストボックスに移動し、**OK**ボタンをクリックします。

low_income

		度数	パーセント	有効パーセント	累積パーセント
有効	.00	77	38.9	38.9	38.9
	1.00	121	61.1	61.1	100.0
	合計	198	100.0	100.0	

図E.21　low_incomeの度数分布表

全体の約６割の人が、収入が地域の平均より低いことがわかります。

8.　収入が地域の平均より高いグループと低いグループを比較し、現在の残高（currbal）に違いがあるか調べます。

7. で求めた変数を使って、２つのグループのcurrbalの平均値を比較します。

☞ **分析**メニューの**平均値の比較**の**グループの平均**を選択します。
currbalを**従属変数**ボックスに、**low_income**（7. で求めた変数）を**独立変数**ボックスに移動し、**OK**ボタンをクリックします。

報告書

現在の残高の平均

low_income	平均値	度数	標準偏差
.00	730.8029	77	927.00134
1.00	833.3249	121	1052.86227
合計	793.4552	198	1004.62397

図E.22　currbalの平均値の比較