IBM SPSS Data Preparation 19

Note: Before using this information and the product it supports, read the general information under Notices第 139 頁.

序

IBM® SPSS® Statistics為分析資料的強大系統。資料準備 的選用性附加模組能提供其他本手冊所說明的分析技術。資料準備 的附加模組必須與 SPSS Statistics Core 系統搭配使用，而且是完全整合到系統中。

關於 SPSS Inc.，是一家 IBM 公司

SPSS Inc.，是一家 IBM 公司，為全球領先的預測分析軟體和解決方案供應商。該公司完整的系列產品 — 資料收集、統計量、模型製造與部署 — 捕捉人們的態度和意見，預測客戶未來的互動結果，然後將分析融入業務程序，以依照所得見解採取行動。SPSS Inc. 解決方案藉由著重於收斂性分析、IT 架構和業務程序，以達成整個組織相互關聯的業務目標。全球商業、政府和學界客戶均仰賴 SPSS Inc. 技術為競爭優勢，以吸引、留住和增加客戶人數，同時減少欺詐並降低風險。SPSS Inc. 在 2009 年 10 月由 IBM 收購。如需詳細資訊，請造訪 http://www.spss.com。

技術支援

技術支援可提供客戶維護的服務。客戶可以電洽技術支援以取得 SPSS Inc. 產品在使用上的協助，或是支援硬體環境的安裝說明。如果要聯絡技術支援，請參閱 SPSS Inc. 網站（網址是 http://support.spss.com），或是透過網站（網址是 http://support.spss.com/default.asp?refpage=contactus.asp）尋找當地的辦事處。請求協助時，請準備好的您個人、組織和支援合約的相關資訊。

客戶服務

如果您對於自己的貨品或帳號有任何疑問，請聯絡您的當地辦公室，列示於網站上：http://www.spss.com/worldwide。請備妥您的序號以供識別。

訓練研討會

SPSS Inc. 同時提供公開與線上訓練研討會。所有的研討會皆以傳達工作群為其特色。研討會將定期在各主要城市舉辦。如需有關這些研討會的更多資訊，請聯絡您的當地辦公室，列示於網站上：http://www.spss.com/worldwide。

其他出版品

SPSS Statistics：Guide to Data Analysis（資料分析指南）、SPSS Statistics：Statistical Procedures Companion（統計程序指南）以及 SPSS Statistics：Advanced Statistical Procedures Companion（進階統計程序指南）是由 Marija Norušis 撰寫，

由 Prentice Hall 發行，為推薦的輔助資料。這些出版品涵蓋 SPSS Statistics Base 模組、進階統計量模組和迴歸模組中的統計程序。不論您是資料分析的新手，還是已經準備使用高階應用程式，這些書籍都能幫助您善加利用 IBM® SPSS® Statistics 系列產品中的功能。如需其他資訊（包括出版品內容和章節樣本），請參閱作者的網站： http://www.norusis.com

內容

附錄

部 I:
使用手冊

章節
1

資料準備簡介

隨著運算系統功能提升，對資料的需求也成比例地上升，導致愈來愈多資料收集—、更多觀察值、更多變數及更多資料輸入錯誤。這些錯誤是預測模型預測值的禍根，這些預測是倉儲的資料最終目標，所以您需要維持資料的「乾淨」。然而，倉儲的資料數量已經無法以手動驗證觀察值，所以執行自動化驗證資料程序是很重要的。

「資料準備」附加模組可讓您識別您的作用中資料集內異常的觀察值及無效的觀察值、變數及資料值，並準備建模用的資料。

使用資料準備程序

「資料準備」程序的使用取決於您的特定需求。載入您的資料後，一般程序為：

- **中繼資料準備。** 檢視您資料檔中的變數並決定其有效數值、標記及測量水準。識別編碼錯誤但無法分析的變數數值組合。根據這項資訊而定義驗證規則。這可能是一個耗時的工作，但如果您需要定期以類似屬性驗證資料檔，這項努力是值得的。
- **資料驗證。** 執行基本檢查並與已定義的驗證規則比對，以識別無效的觀察值、變數及資料值。發現無效資料時，調查並更正其原因。可能需要進行中繼資料準備中的另一個步驟。
- **模式準備。** 使用自動資料準備以取得可改善建模的原始欄位轉換。識別可能導致許多預測模型問題的潛在統計偏離值。部分偏離值是由尚未識別的無效變數值所導致的。可能需要進行中繼資料準備中的另一個步驟。

一旦資料檔「乾淨」，您就可以從其他附加模組建立模式。

章節
2

驗證規則

驗證觀察值是否有效的規則。驗證規則有兩種：

- **單一變數規則。** 單一變數規則包含一組套用至單一變數的固定檢查項目，如檢查數值是否超出範圍等。對於單一變數規則，有效值可表示為數值範圍，或是可接受數值清單。
- **交叉變數規則。** 交叉變數規則是使用者定義的規則，可套用至單一變數或變數組合。交叉變數規則可由標示無效數值的邏輯運算式定義。

驗證規則會儲存在您資料檔案的資料目錄。這可讓您指定規則並再次使用之。

載入預先定義的驗證規則

您可從安裝中包含的外部資料檔案載入預先定義的規則，快速取得一組已可使用的驗證規則。

載入預先定義的驗證規則

▶ 從功能表選擇：

資料 > 驗證 > 載入預先定義的規則...

圖表 2-1
載入預先定義的驗證規則

請注意，此程序會刪除作用中資料集中任何現有單一變數規則。

您可改用「複製資料性質精靈」，從任何資料檔案載入規則。

定義驗證規則

「定義驗證規則」對話方塊可讓您建立並檢視單一變數與交叉變數驗證規則。

建立並檢視驗證規則

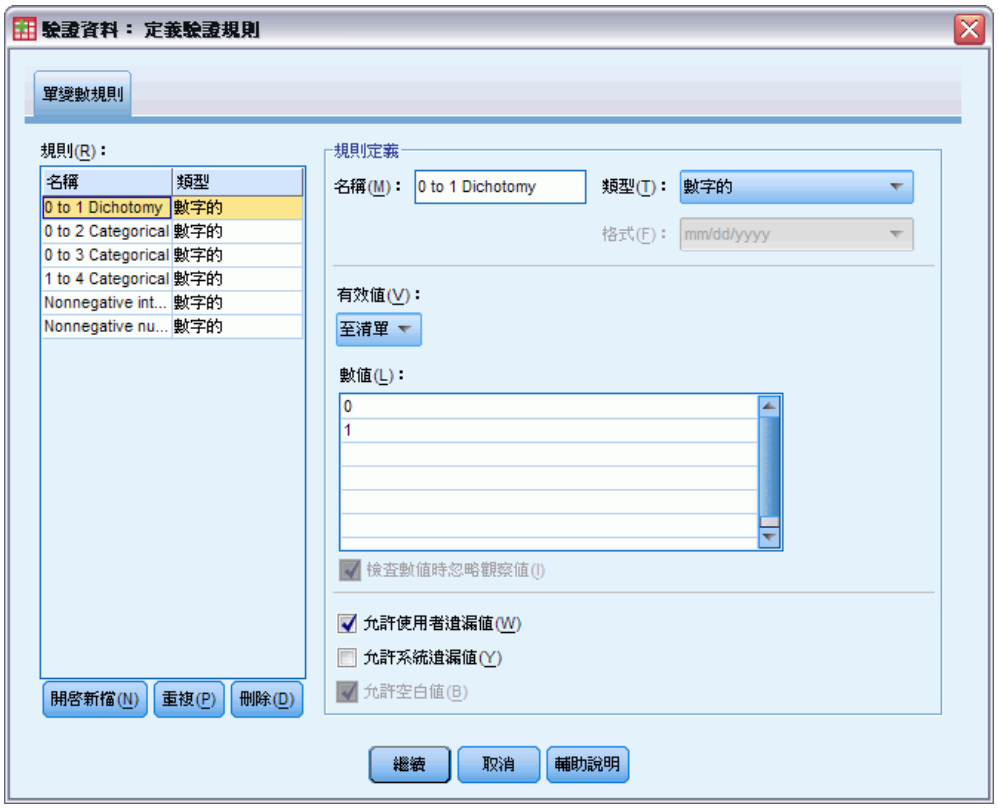
- 從功能表選擇：
  資料 > 驗證 > 定義規則...

  此對話方塊中集合了從資料目錄中讀取到的單一變數與交叉變數驗證規則。若無規則，會自動建立新預留位置規則，讓您可進行修改以符合需求。

- 在「單一變數規則」和「交叉變數規則」索引標籤中選擇個別的規則，來進行檢視並修改性質。

定義單一變數規則

圖表 2-2
「定義驗證規則」對話方塊，「單一變數規則」索引標籤



「單一變數規則」索引標籤可讓您建立、檢視和修改單一變數驗證規則。

規則。 清單依要套用規則的變數名稱與類型，顯示單一變數驗證規則。開啟對話方塊時，會顯示資料目錄中定義的規則，或目前未定義規則時，會顯示名為「單一變數規則 1」的預留位置規則。「規則」清單下會有下列按鈕：

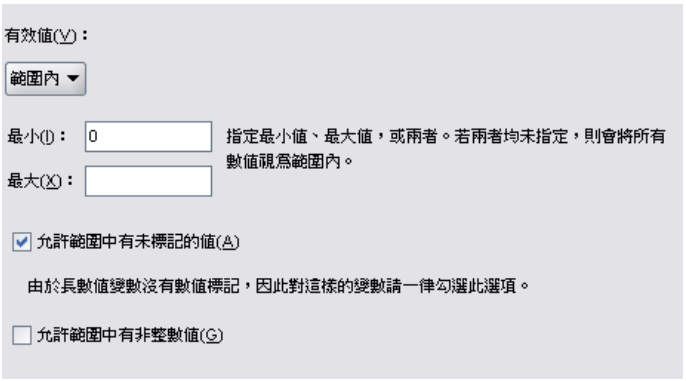
- **開啟新檔。** 將新項目新增至「規則」清單下。此所選規則會被指定名稱「SingleVarRule n」，其中 n 是一整數，這樣單一變數與交叉變數的新規則名稱都會是唯一的。
- **重複。** 將所選規則的副本新增至「規則」清單下。該規則名稱會被調整，以讓每個單一變數或交叉變數規則名稱均為唯一的。例如，若您複製「SingleVarRule 1」，則第一個規則副本的名稱會是「副本 SingleVarRule 1」，第二個副本名稱為「副本 (2) SingleVarRule 1」，以此類推。
- **刪除。** 刪除選定的規則。

規則定義。 這些控制項可讓您檢視並設定所選規則的性質。

- **名稱。** 單一變數與交叉變數規則名稱均必須為唯一的。
- **類型。** 這是要套用規則的變數類型。請從**數值**、**字串**和**日期**之間選擇。
- **格式。** 這可讓您選擇要套用至日期變數的日期格式規則。
- **有效值。** 您可指定數值範圍或清單為有效值。

範圍定義控制項可讓您指定有效範圍。超出範圍的數值會標示為無效。

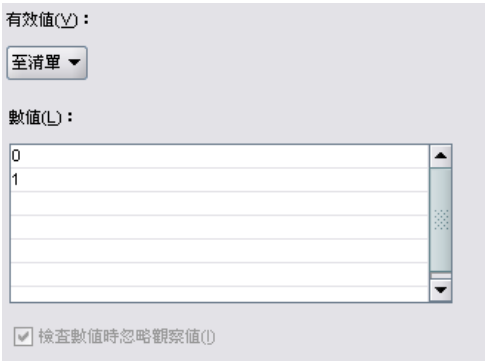
圖表 2-3
單一變數規則：範圍定義



若要定義範圍，請輸入最小值或最大值，或兩者。核取方塊控制項可讓您標示範圍內的未標記或非整數數值。

清單定義控制項可讓您定義有效數值清單。未包含於此清單的數值會被標示為無效。

圖表 2-4
單一變數規則：清單定義

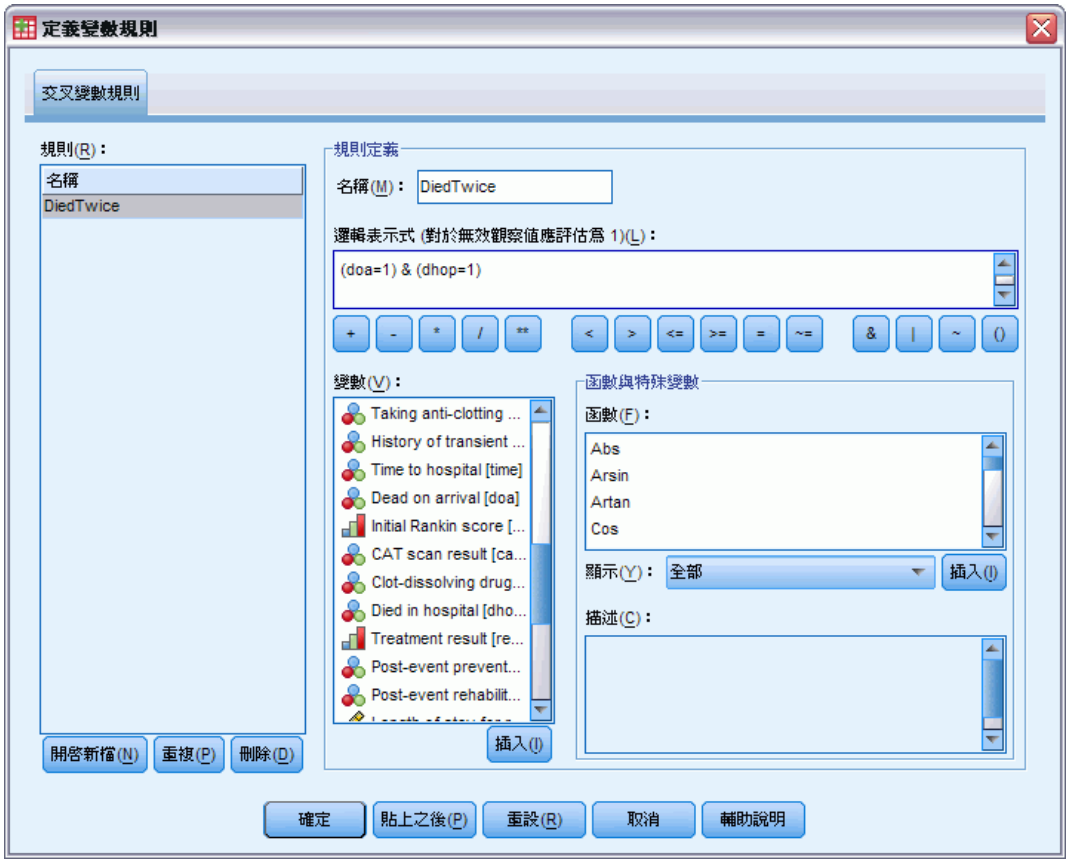


在格線中輸入清單數值。以可接受值清單檢查字串資料值時，核取方塊會判斷觀察值是否有效。

- **允許使用者遺漏值。** 控制是否要將使用者遺漏值標示為無效。
- **允許系統遺漏值。** 控制是否要將系統遺漏值標示為無效。這不會套用至字串規則項目。
- **允許空白值。** 控制是否將空白（表示完全空白）字串標示為無效。這不會套用至非字串規則項目。

定義交叉變數規則

圖表 2-5
「定義驗證規則」對話方塊，「交叉變數規則」索引標籤



「交叉變數規則」標籤可讓您建立、檢視和修改交叉變數驗證規則。

規則。 清單會依名稱顯示交叉變數驗證規則。開啟對話方塊時，會顯示名為「CrossVarRule 1」的預留位置規則。「規則」清單下會有下列按鈕：

- **開啟新檔。** 將新項目新增至「規則」清單下。此所選規則會被指定名稱「CrossVarRule n」，其中 n 是一整數，這樣單一變數與交叉變數的新規則名稱都會是唯一的。

- **重複。** 將所選規則的副本新增至「規則」清單下。該規則名稱會被調整，以讓每個單一變數或交叉變數規則名稱均為唯一的。例如，若您複製「CrossVarRule 1」，則第一個規則副本的名稱會是「副本 CrossVarRule 1」，第二個副本名稱為「副本 (2) CrossVarRule 1」，以此類推。
- **刪除。** 刪除選定的規則。

規則定義。 這些控制項可讓您檢視並設定所選規則的性質。

- **名稱。** 單一變數與交叉變數規則名稱均必須為唯一的。
- **邏輯運算式。** 事實上，這是規則定義。您應編碼運算式，將無效觀察值評估為 1。

建立運算式

▶ 若要建立運算式，請將組成成份貼入「運算式」欄位，或者直接輸入「運算式」欄位中。

- 您可從「函數」群組清單選擇群組來貼上函數或常用的系統變數，並按兩下「函數與特殊變數」清單中的函數或變數（或選擇函數或變數並按一下**插入**）。填入標有問號的所有參數（僅適用函數）。標示為**全部**的函數群組會列出所有可用函數與系統變數。對話方塊的保留區域會顯示簡要的描述，說明目前選取的函數或變數。
- 字串常數必須括在引號或撇號中。
- 如果數值中包含有小數點，必須使用句點 (.) 作為小數點符號。

章節
3

驗證資料

「驗證資料」對話方塊可以讓您識別可疑的和無效的觀察值、變數，以及在作用中資料集中的資料值。

範例。 資料分析師必須將每月客戶滿意度報告提供給她的客戶。資料分析師必須針對每個月所收到的資料進行品質檢查，包括，不完整客戶 ID、超出範圍的變數數值，以及經常輸入錯誤之變數數值的組合。「驗證資料」對話方塊可以讓資料分析師設定能唯一識別顧客的變數、定義有效變數範圍的單一變數規則，以及定義交叉變數規則以找到不可能的組合。這個程序會傳回有問題之觀察值與變數的報告。此外，還會傳回每個月含有相同資料元素的資料，因此分析師可以將規則套用到下個月的新資料檔案中。

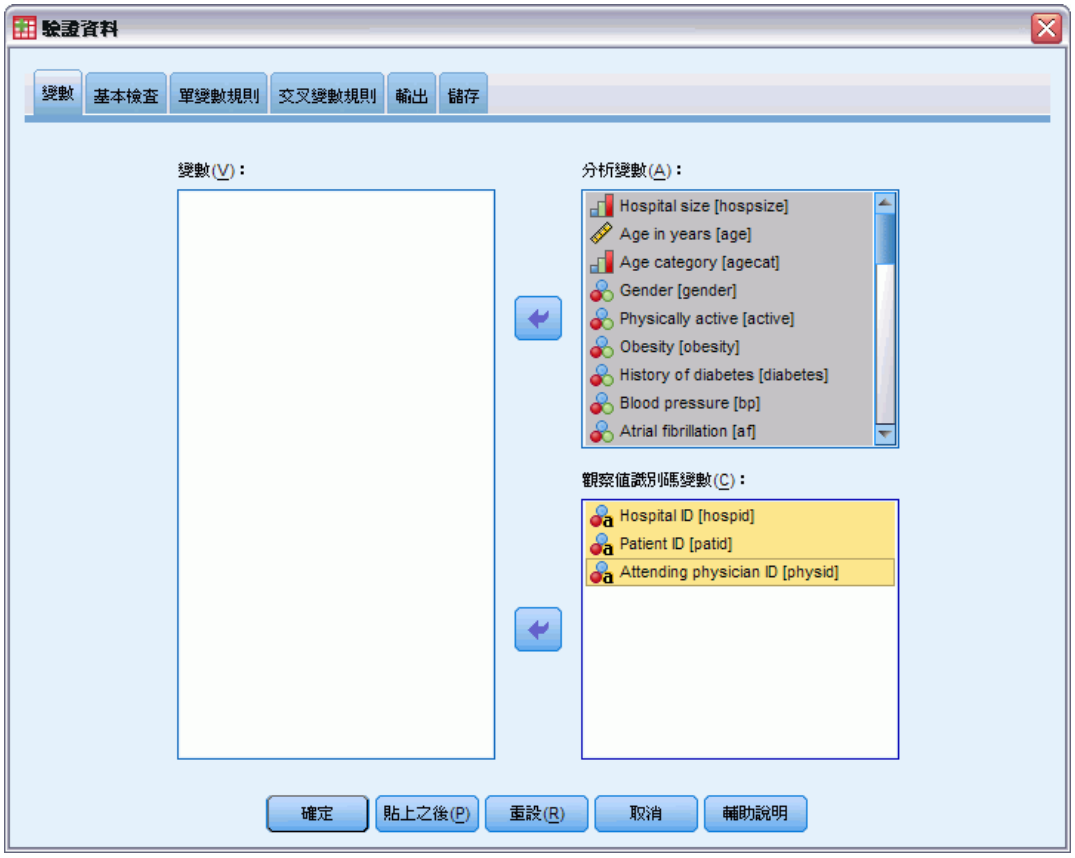
統計量。 這個程序會產生變數、觀察值，和沒有通過各項檢查的資料數值清單、單一變數和交叉變數違規次數，以及有關分析變數的簡單說明摘要。

加權值。 這個程序會忽略加權變數規格，並且將它當成任何其他的分析變數處理。

若要驗證資料

- 從功能表選擇：
  資料 > 驗證(V) > 驗證資料(V)...

圖表 3-1
「驗證資料」對話方塊，「變數」索引標籤



▶ 根據基本變數檢查或單一變數驗證規則來選擇一個或多個用來驗證的分析變數。

您也可以：

▶ 按一下「**交叉變數規則**」索引標籤，並且套用一個或多個交叉變數規則。

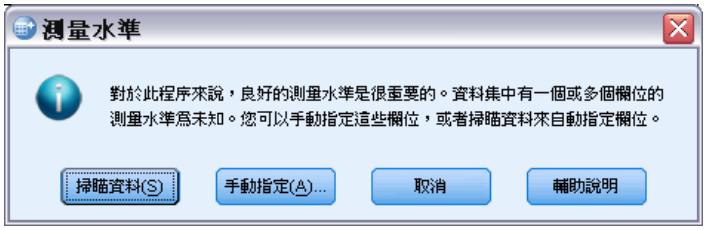
您可以：

- 選擇一個或多個觀察值辨識變數，以檢查重複或不完整 ID。觀察值 ID 變數也可以用來標示觀察值輸出。如果指定兩個（或以上）個觀察值 ID 變數時，會將這些數值的組合當做觀察值識別碼來處理。

具有未知測量水準的欄位

若在資料集中出現一或多個未知的變數（欄位）測量水準，就會顯示「測量水準」警示。由於測量水準會影響此程序的結果計算，因此所有變數皆必須具有已定義的測量水準。

圖表 3-2
測量水準警示



- **掃描資料。** 讀取作用中資料集的資料，並且針對目前具有未知測量水準的任何欄位指派預設的測量水準。若為大型資料集，則讀取時可能需要一些時間。
- **手動指派。** 開啟對話方塊，以列出具有未知測量水準所有欄位。您可以使用此對話方塊，來指派上述欄位的測量水準。您也可以在「資料編輯程式」的「變數檢視」中指派測量水準。

由於測量水準是此程序的重要項目，因此您在所有欄位皆擁有已定義的測量水準之前，無法存取對話方塊來執行此程序。

驗證資料基本檢查

圖表 3-3
「驗證資料」對話方塊，「基本檢查」索引標籤

「基本檢查」索引標籤可以讓您選擇分析變數、觀察值識別碼，以及整個觀察值的基本檢查。

分析變數。 如果已經選擇「變數」索引標籤上的任何分析變數，您就可以選擇以下任何有效性的檢查。核取方塊可以讓您核取或取消勾選。

- **遺漏值的最大百分比。** 報告中會分析遺漏值百分比大於指定值的變數。指定值必需是小於或等於 100 的正數。
- **單一類別中觀察值的最大百分比。** 如果有任何分析變數是類別的，則這個選項會報告代表單一非遺漏類別之觀察值百分比大於指定值的類別分析變數。指定值必須為小於或等於 100 的正數。百分比是以含有非遺漏值變數的觀察值為根據。
- **個數 1 之類別的最大百分比。** 如果有任何分析變數是類別的，則這個選項會報告變數類別百分比中只有一個觀察值大於指定值的類別分析變數。指定值必需是小於或等於 100 的正數。

- **最小變異係數。** 如果有任何分析變數是尺度，則這個選項會報告變異係數絕對值小於指定值的尺度分析變數。這個選項只套用於其平均數不是零的變數。指定值必須為非負數值。指定 0 會關閉變異係數檢查。
- **最小標準差。** 如果有任何分析變數是尺度，則這個選項會報告標準差小於指定值的尺度分析變數。指定值必須為非負數值。指定 0 會關閉標準差檢查。

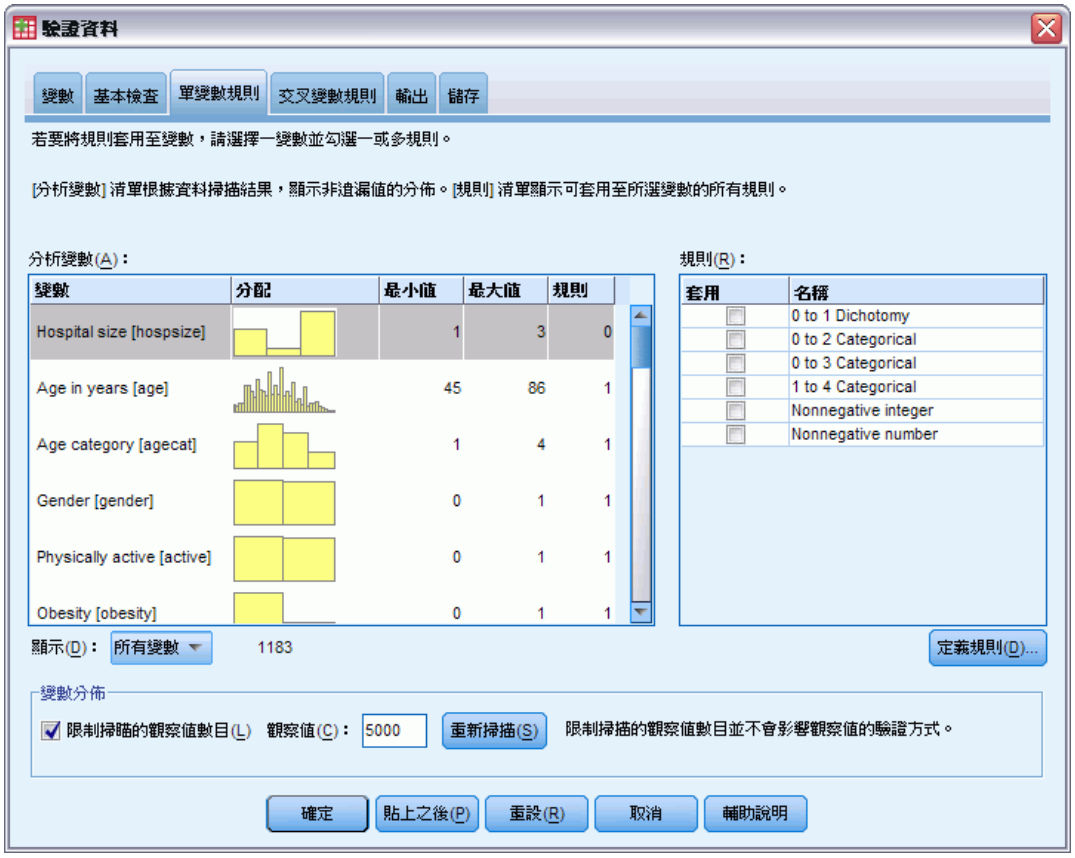
觀察值識別碼。 如果已經選擇「變數」索引標籤上的任何觀察值識別碼變數，您就可以選擇以下任何其有效性的檢查。

- **標示不完整 ID。** 這個選項會報告含有不完整觀察值識別碼的觀察值。如果是特定的觀察值，當任何 ID 變數的數值為空白或遺漏時，則視識別碼為不完整。
- **標示重複 ID。** 這個選項會報告含有重複觀察值識別碼的觀察值。會將不完整識別碼從可能的重複值組中排除。

標示空白觀察值。 這個選項會報告所有變數為空白的觀察值。為了識別空白觀察值，您可以選擇使用所有檔案中變數（任何 ID 變數除外），或只選擇使用在「變數」索引標籤上所定義的分析變數。

驗證資料單一變數規則：

圖表 3-4
「驗證資料」對話方塊，「單一變數規則」索引標籤



「單一變數規則」索引標籤會顯示可用的單一變數驗證規則，而且可以讓您套用到分析變數中。若要定義其他單一變數規則，請按一下「**定義規則**」。

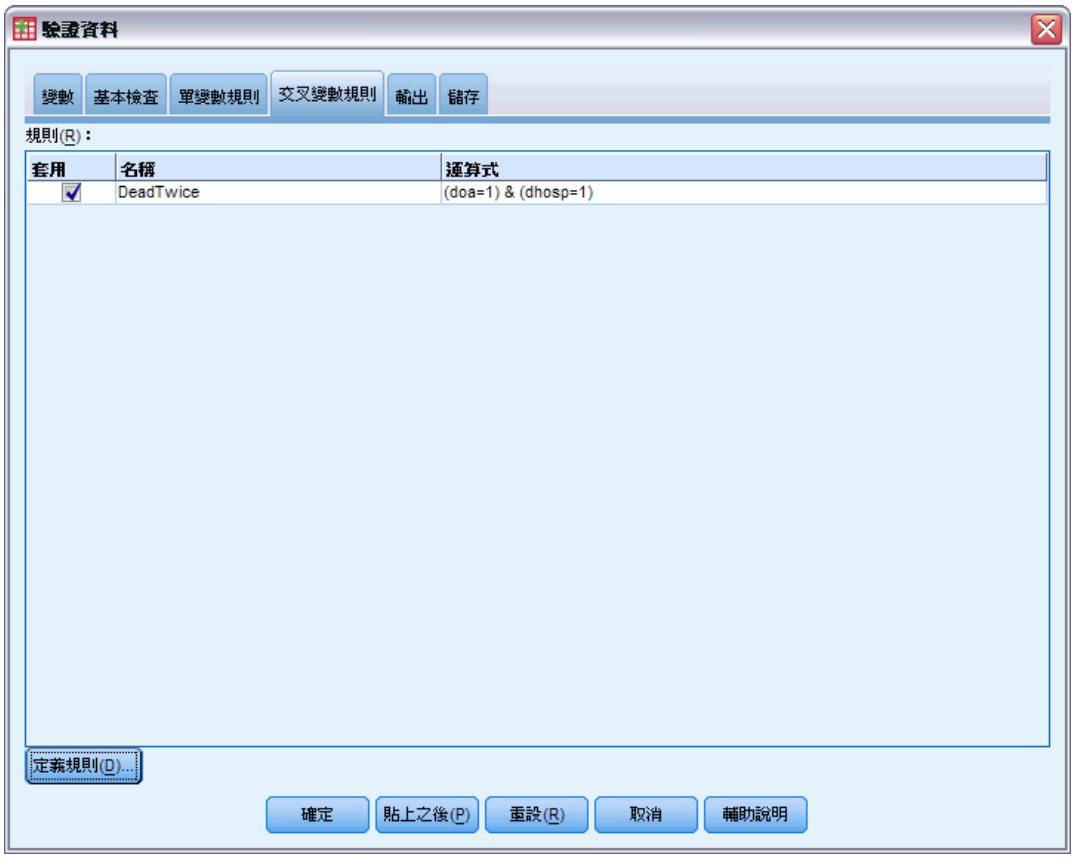
分析變數。 這個清單會顯示分析變數、摘要其分配狀態，並且顯示套用到各個變數的規則數目。請注意，摘要中不包含使用者和系統遺漏值。「顯示」下拉式清單會控制要顯示哪一個變數，您可以從「**所有變數**」、「**數值變數**」、「**字串變數**」和「**日期變數**」中選取。

規則。 若要套用規則到分析變數中，請選擇一個或多個變數，並且核取您要套用在「規則」清單中的所有規則。「規則」清單只會顯示已選擇之分析變數所適用的規則。例如，如果選擇數值分析變數時，就只會顯示數字規則，如果選擇字串變數時，則只會顯示字串規則。如果都沒有選擇分析變數，或這些變數含有混合資料類型，則不會顯示規則。

變數分配。「分析變數」清單中的所顯示分配摘要，是以所有觀察值為根據，或是以前 n 個觀察值的掃描為根據，如「觀察值」文字方塊中所指定。按一下「**重新掃描**」，更新分配摘要。

驗證資料交叉變數規則

圖表 3-5
「驗證資料」對話方塊，「交叉變數規則」索引標籤



「交叉變數規則」索引標籤會顯示可用的交叉變數規則，而且可以讓您套用到您的資料中。若要定義其他交叉變數規則，請按一下「**定義規則**」。

驗證資料輸出

圖表 3-6
「驗證資料」對話方塊，「輸出」索引標籤

逐觀察值報告。 如果您已經套用任何單一變數或交叉變數驗證規則，您可以要求一份列出個別觀察值驗證規則違規的報告。

- **最小違規數。** 這個選項會指定要包含在報告中的觀察值所需之最小規則違規數。指定一個正整數。
- **最大觀察值個數。** 這個報告會指定包含在觀察值報告中的最大觀察值個數。指定小於或等於 1000 的正整數。

單一變數驗證規則。 如果您已經套用任何單一變數驗證規則，您就可以選擇顯示結果的方法，或是是否要顯示結果。

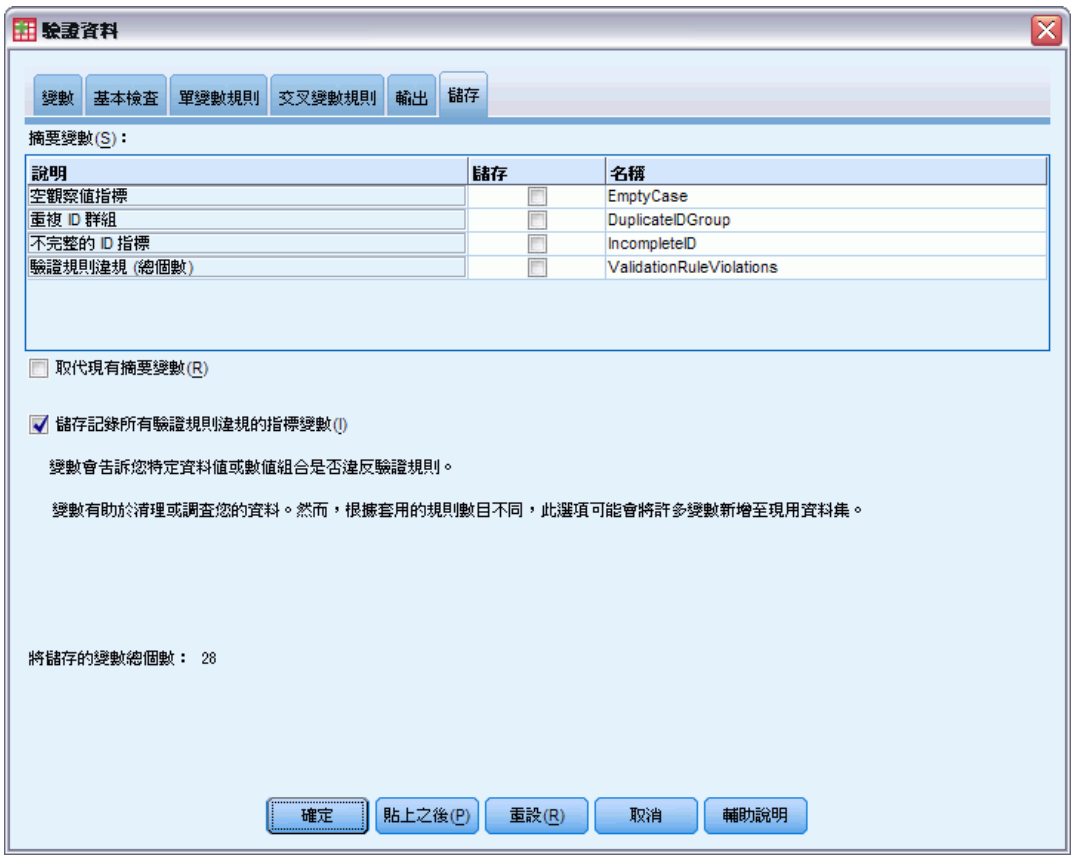
- **根據分析變數摘要違規。** 如果是各個分析變數，這個選項會顯示所有被違反的單一變數驗證規則，以及所違反的每一規則之數值的個數。也會報告各個變數之單一變數規則違規的總數。
- **根據規則摘要違規。** 如果是各個單一變數驗證規則，這個選項會報告違反規則的變數，以及每一個變數之無效值的個數。也會報告所有變數數值違規的總數。

顯示描述性統計量。 這個選項可以讓您要求分析變數的描述性統計量。會為各個類別變數產生次數表。會為尺度變數產生包含平均數、標準差、最小值，和最大值的摘要統計量表。

移動含有驗證規則違規的觀察值。 這個選項會將含有單一變數或交叉變數規則違規的觀察值，移到作用中資料集的頂端以方便仔細觀察。

驗證資料儲存

圖表 3-7
「驗證資料」對話方塊，「儲存」索引標籤



「儲存」索引標籤可以讓您儲存將規則違規紀錄到作用中資料集的變數。

摘要變數。 這些是可以儲存的個別變數。核取一個方塊以儲存變數。會提供變數的預設名稱，您可以進行編輯。

- **空白觀察值指標。** 空白觀察值會指定為數值 1，所有其他觀察值則編碼為 0。變數的數值會反應「基本檢查」索引標籤上所指定的範圍。
- **重複 ID 群組。** 含有相同觀察值識別碼的觀察值（而不是含有不完整識別碼的觀察值），會指定相同的組別號碼。會將含有唯一或不完整識別碼的觀察值編碼為 0。
- **不完整的 ID 指標。** 含有空白或不完整觀察值識別碼的觀察值會指定為數值 1。其他觀察值則編碼為 0。
- **驗證規則違規。** 這是單一變數和交叉變數驗證規則違規的觀察值總數。

取代現有的摘要變數。 儲存於資料檔案的變數名稱必須是唯一的，否則會取代具有相同名稱的變數。

儲存指標變數。 這個選項可以讓您儲存驗證規則違規的完整記錄。各個變數都會對應到一個驗證規則的應用程式，而且如果觀察值違反規則時就會含有數值 1，如果沒有違反規則，則會含有數值 0。

章節
4

自動資料準備

準備資料以供分析是任何專案中最重要的步驟之一—也是傳統上最耗時的步驟之一。「自動資料準備」(ADP) 可為您處理工作、分析您的資料並識別修正、篩選出有問題或可能無用的欄位、在適當時衍生新屬性，以及透過智慧型篩選技術增進效能。您可以全**自動**方式使用演算法，以允許其選擇並套用修正，或以**互動**方式使用演算法，以在進行變更前先行預覽，然後視需要接受或拒絕變更。

使用 ADP 可讓您快速、輕鬆地準備資料以建立模式，不需事先了解統計相關概念。模式將可更快地建立並進行資料評分，此外，使用 ADP 可提高自動建立模式程序。

注意：ADP 準備要進行分析的欄位時，會建立包含調整或轉換的新欄位，而非取代舊欄位現有的值和性質。舊欄位不會用於進一步分析；其角色會設定為「無」。 此外亦請注意，系統不會將任何使用者遺漏值資訊轉換至這些新建立的欄位，若在新欄位中存有任何遺漏值，則會歸為系統遺漏值。

範例。 某資源有限的保險公司，打算調查屋主的保險理賠，希望建立標示可疑潛在詐欺理賠的模式。建立模式之前，他們將使用自動資料準備來準備建模用的資料。由於他們希望在套用轉換前檢閱提議的轉換，因此會在互動式模式使用自動資料準備。

某汽車業集團會追蹤各種個人汽車的銷售額。為了能夠識別表現超前與表現不佳的模式，他們希望建立汽車銷售額與汽車特性之間的關係。他們會使用自動的資料準備來準備分析用的資料，以及使用準備「之前」與「之後」的資料建立模式，以瞭解結果有何差異。

圖表 4-1
「自動資料準備目標」索引標籤

您的目標是什麼？ 自動資料準備會建議資料準備步驟，這些步驟將影響其他演算法建立模式的速度，並提升這些模式的預測能力。其中包含轉換、建立和選取功能。亦可轉換目標。您可以指定資料準備步驟遵循的模式建立優先順序。

- **權衡速度與準確度。** 此選項準備資料時，會兼顧模式建立演算法處理資料的速度，以及預測的準確度。
- **最佳化速度。** 此選項準備資料時，會優先考慮模式建立演算法處理資料的速度。當您正在處理非常大型的資料集或想快速找到答案時，請選取此選項。
- **最佳化準確度。** 此選項準備資料時，會優先考慮模式建立演算法所產生預測的準確度。
- **自訂分析。** 當您想在「設定」索引標籤中手動變更演算法時，請選取此選項。請注意，若您之後對「設定」索引標籤中的選項進行變更，但該變更與任一項目標不符時，會自動選取此設定。

取得自動資料準備

從功能表選擇：
轉換(T) > 準備建模用的資料 > 自動式(A)...

▶ 按一下「**執行**」。

您可以：

- 在「目標」索引標籤上指定目標。
- 在「欄位」索引標籤上指定欄位指派。
- 在「設定」索引標籤上指定匯出設定。

取得互動式資料準備

從功能表選擇：
轉換(T) > 準備建模用的資料 > 互動式(N)...

▶ 在對話方塊上方的工具列中，按一下「**分析**」。

▶ 按一下「分析」索引標籤並檢視建議的資料準備步驟。

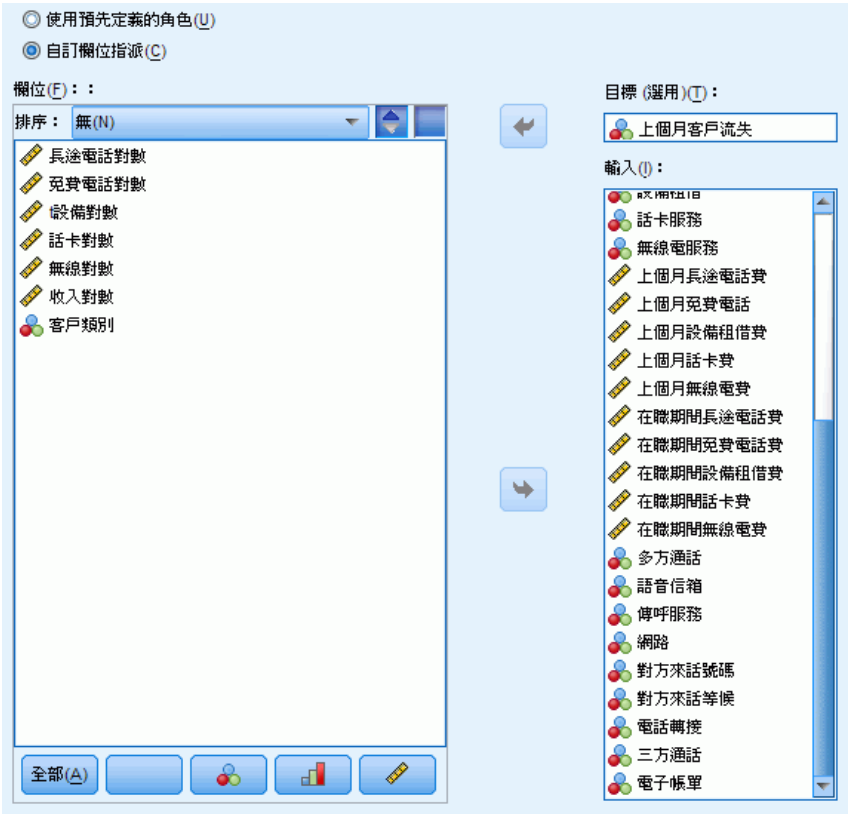
▶ 如果滿足您的需求，請按一下「**執行**」。否則，請按一下「**清除分析**」，視需要變更任何設定，然後按一下「**分析**」。

您可以：

- 在「目標」索引標籤上指定目標。
- 在「欄位」索引標籤上指定欄位指派。
- 在「設定」索引標籤上指定匯出設定。
- 按一下「**儲存 XML**」，將建議的資料準備步驟儲存到 XML 檔案。

欄位索引標籤

圖表 4-2
「自動資料準備欄位」索引標籤



「欄位」索引標籤指定應準備哪些欄位以進一步分析。

使用預先定義的角色。 此選項使用現有的欄位資訊。若有一個欄位含有「目標」角色，則會將其當做目標；否則將不會有目標。含有預先定義角色「輸入」的所有欄位都將做為輸入。至少需要一個輸入欄位。

使用自訂欄位指派。 從欄位的預設清單移動欄位來覆寫欄位角色時，對話方塊將自動切換至此選項。進行自訂欄位指派時，請指定下列欄位：

- **目標（選用）。**若您計畫建立需要目標的模式，請選取目標欄位。這與將欄位角色設定為「目標」相同。
- **輸入。**選取一或多個輸入欄位。這與將欄位角色設定為「輸入」相同。

設定索引標籤

「設定」索引標籤含有多種設定群組，可讓您修改以微調演算法處理資料的方式。若您對預設設定所做的任何變更與其他目標不符，「目標」索引標籤會自動更新為選取「自訂分析」選項。

準備日期與時間

圖表 4-3
自動資料準備的「準備日期與時間」設定

許多模式建立演算法均無法直接處理日期與時間詳細資料；這些設定可讓您衍生新的期間資料，以做為您現有資料中日期和時間的模式輸入。包含日期與時間的欄位必須預先定義日期或時間儲存類型。原始的日期與時間欄位在自動資料準備之後將不建議做為模式輸入。

準備建模的日期與時間。 取消選取此選項會停用全部其他「準備日期與時間」控制項，同時維持選擇。

計算至參考日期需經過的時間。 這會產生自各包含日期變數的參考日期至今的年/月/天數。

- **參考日期。**指定輸入資料的日期資訊中，做為計算持續期間起始日的日期。選取「今天日期」表示執行 ADP 時，永遠會使用目前的系統日期。若要使用特定日期，請選取「固定日期」並輸入必要的日期。
- **日期持續期間的單位。**指定 ADP 應自動決定日期持續期間的單位，或從「年數」、「月」或「天數」的「固定單位」中選取。

計算至參考時間需經過的時間。 這會產生自各包含時間變數的參考時間至今的小時/分鐘/秒數。

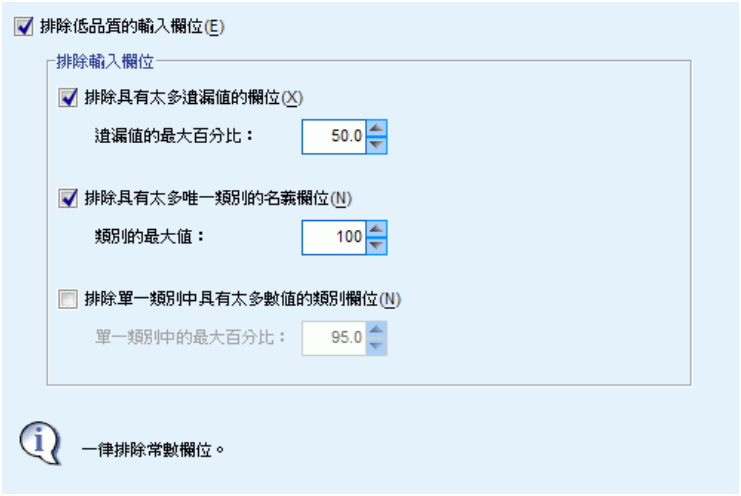
- **參考時間。**指定輸入資料的時間資訊中，做為計算持續期間起始時間的時間。選取「目前時間」表示執行 ADP 時，永遠會使用目前的系統時間。若要使用特定時間，請選取「固定時間」並輸入必要的詳細資料。
- **時間持續期間的單位。**指定 ADP 應自動決定時間持續期間的單位，或從「時數」、「分鐘數」或「秒數」的「固定單位」中選取。

萃取循環時間元素。使用這些設定將單一日期或時間欄位分割為一或多個欄位。例如，若您選取這三個日期的核取方塊，輸入日期欄位 “1954-05-23” 會分割為三個欄位：1954、5 和 23，且會分別使用「固定名稱」面板中定義的字尾，並且會忽略原始日期。

- **從日期萃取。**對於任何日期輸入，指定您要萃取年、月、日或任何組合。
- **從時間萃取。**對於任何時間輸入，指定您要萃取小時、分鐘、秒或任何組合。

排除欄位

圖表 4-4
自動資料準備的「排除欄位」設定



品質不佳的資料會影響預測的準確度；因此，您可以指定可接受的輸入等級品質功能。所有常數欄位或含有 100% 遺漏值的欄位都會自動被排除。

排除低品質的輸入欄位。 取消選取此選項會停用全部其他「排除欄位」控制項，同時維持選擇。

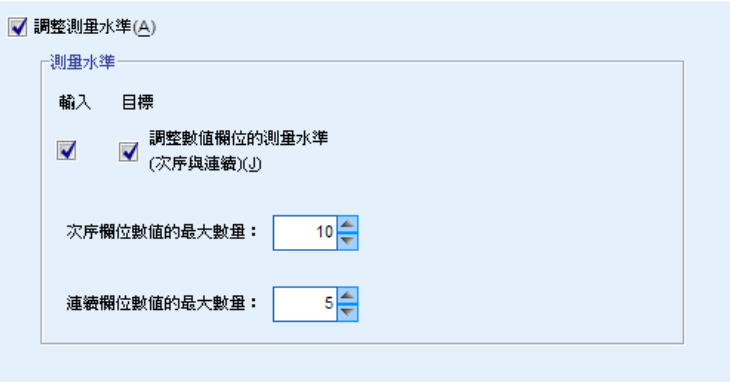
排除具有太多遺漏值的欄位。超過指定遺漏值百分比的欄位會被移除，不執行進一步分析。即使指定大於或等於 0（等於取消選取此選項），而且小於或等於 100 的數值，所有含有遺漏值的欄位還是會遭自動排除。預設值是 50。

排除具有太多唯一類別的名義欄位。超過指定類別數目的名義欄位會被移除，不執行進一步分析。指定一個正整數。預設值是 100。這對自動從建模移除包含記錄唯一資訊（例如 ID、位址或名稱）的欄位很實用。

排除單一類別中具有太多數值的類別欄位。含有超過指定記錄百分比之類別的次序和名義欄位會被移除，不執行進一步分析。即使指定大於或等於 0（等於取消選取此選項），而且小於或等於 100 的數值，常數欄位還是會遭自動排除。預設值是 95。

調整測量

圖表 4-5
自動資料準備的「調整測量」設定



調整測量水準。 取消選取此選項會停用全部其他「調整測量」控制項，同時維持選擇。

測量水準。指定含有「太少」值之連續欄位的測量水準是否可調整為次序，以及含有「太多」值之次序欄位的測量水準是否可調整為連續。

- **次序欄位數值的最大數量。**超過指定類別數目的次序欄位會重新分配為連續欄位。指定一個正整數。預設值是 10。此值必須大於或等於連續欄位值的最小數目。
- **連續欄位數值的最大數量。**少於指定唯一值數目的連續欄位會重新分配為次序欄位。指定一個正整數。預設值是 5。此值必須小於或等於次序欄位值的最小數目。

改進資料品質

圖表 4-6
自動資料準備的「改進資料品質」設定

準備要改進資料品質的欄位。 取消選取此選項會停用全部其他「改進資料品質」控制項，同時維持選擇。

偏離值處理。指定是否置換輸入與目標的偏離值；若是如此，則指定偏離值分割條件（在標準差中測量）以及置換偏離值的方法。偏離值可透過刪除（設定為分割值）或將其設定為遺漏值來置換。任何設為遺漏值的偏離值，都會依循在下面選取的遺漏值處理設定。

置換遺漏值。指定是否置換連續、名義或次序欄位的遺漏值。

重新排序名義欄位。選取此項以重新編碼名義（已設定）欄位的值（從最小（最不常出現）到最大（最常出現）類別。新欄位數值會以 0 開頭，做為次數最少的類別。請注意，即使原始欄位為字串，新欄位仍會是數字。例如，如果名義欄位的資料數值為「A」、「A」、「A」、「B」、「C」、「C」，則自動的資料準備會重新編碼「B」為 0、「C」為 1，而「A」為 2。

重新調整欄位大小

圖表 4-7
自動資料準備的「重新調整欄位大小」設定

重新調整欄位大小。 取消選取此選項會停用全部其他「重新調整欄位大小」控制項，同時維持選擇。

分析加權。 此變數包含分析（迴歸或取樣）加權。分析加權是用來說明目標欄位不同等級間的變異數差異。選取連續欄位。

連續輸入欄位。這會使用 **z-分數轉換**或**最小/最大值轉換**來常態化連續輸入欄位。當您在「選取與建立」設定中選取「**執行功能建構**」時，重新調整輸入大小特別有用。

- **z-分數轉換。**此欄位使用觀察的平均數和標準差做為母群參數估計值以進行標準化，接著 z 分數會對應至具有指定之「**最終平均數**」和「**最終標準差**」的對應常態分配值。為「**最終平均數**」指定一個數目，並為「**最終標準差**」指定一個正數。預設值為 0 和 1，分別對應至標準化的重新調整方法。
- **最小/最大值轉換。**此欄位使用觀察的最小值和最大值做為母群參數估計值，對應至具有指定之「**最小值**」和「**最大值**」的對應均勻分配值。指定「**最大值**」大於「**最小值**」的數目。

連續目標。這會使用 Box-Cox 轉換將連續目標轉換為含有接近常態分配（具有指定之「**最終平均數**」和「**最終標準差**」）的欄位。為「**最終平均數**」指定一個數目，並為「**最終標準差**」指定一個正數。預設值分別是 0 和 1。

注意：若某個目標已被 ADP 轉換，後續的模式會使用轉換後的目標分數和單位建立。為解讀和使用結果，您必須將預測值轉換回原始尺度。

轉換欄位

圖表 4-8
自動資料準備的「轉換欄位」設定

若要改善資料的預測能力，您可以轉換輸入欄位。

轉換要進行建模的欄位。 取消選取此選項會停用全部其他「轉換欄位」控制項，同時維持選擇。

類別輸入欄位

- **合併稀疏類別，以最大化與目標之間的關聯。**選取此項以透過減少要處理的目標相關欄位數目，以建立較精簡的模式。相同的類別是根據輸入和目標之間的關係來識別。沒有顯著差異（即 p 值大於指定值）的類別都會被合併。指定大於 0 且小於或等於 1 的值。若所有類別合併為一個，則會從進一步的分析中排除原始和衍生的欄位版本，因為它們沒有當作預測值的值。
- **若無目標，則根據個數合併稀疏類別。**若資料集沒有目標，您可以選擇合併次序與名義欄位的稀疏類別。相同次數方法用於合併含有少於記錄總數之指定最小百分比的類別。指定大於或等於 0 且小於或等於 100 的值。預設值是 10。當沒有包含少於指定觀察值最小百分比的類別時或只有兩個類別時。合併就會停止。

連續輸入欄位。若資料集包含類別目標，您可以極大關聯來 bin 處理連續輸入以改善處理效能。Bin 會根據「同質子集」的性質建立，這是透過使用以指定的 p 值做為關鍵值之 alpha 的 Scheffe 方法所識別，以判斷同質子集。指定大於 0 且小於或等於 1 的一個數值。預設值是 0.05。若 binning 作業會導致特定欄位有一個 bin，則會排除次序和經過 bin 處理之版本的欄位，因為它們沒有做為預測值的值。

注意：ADP 中的 binning 和最適 binning 不同。最適 binning 使用熵資訊來將連續欄位轉換為類別欄位；這需要排序資料並將其全部儲存在記憶體中。ADP 使用同質子集來 bin 處理連續欄位，表示 ADP binning 不需要排序資料，也不會將所有資料儲存在記憶體中。使用同質子集方法來 bin 處理連續欄位表示，經過 bin 處理後的類別數目，永遠會小於或等於目標中的類別數目。

選取與建立

圖表 4-9
自動資料準備中的「選取與建立」設定

為提升資料的預測能力，您可以根據現有的欄位來建立新欄位。

執行功能選擇。若連續輸入與目標的相關性 p 值大於指定的 p 值，就會從分析中移除連續輸入。

執行功能建構。選取此選項，從數個現有功能的組合衍生新功能。舊功能不會用於進一步分析。此選項只適用於目標是連續或沒有目標的連續輸入功能。

欄位名稱

圖表 4-10
自動資料準備的「名稱欄位」設定

為輕鬆識別新功能和轉換功能，ADP 會建立並套用基本新名稱、字首及字尾。您可以修正這些名稱，以更符合您的需求與您的資料。

已轉換與已建構的欄位。指定要套用至轉換後的目標和輸入欄位的副檔名。

此外，請指定要套用至透過「選取」和「建構」設定建構之任何功能的字首名稱。如此便會透過將數值字尾附加到此字首根名稱的方式來建立新名稱。數字的格式會根據衍生多少新功能而定，例如：

- 1-9 個建構的功能將命名為：功能 1 到 功能 9。
- 10-99 個建構的功能將命名為：功能 01 到 功能 99。
- 100-999 個建構的功能將命名為：功能 001 到 999，依此類推。

這可確保無論有多少個功能，建構的功能將依據合理的順序排序。

從日期與時間計算的持續時間。指定副檔名以套用至從日期與時間計算的持續時間。

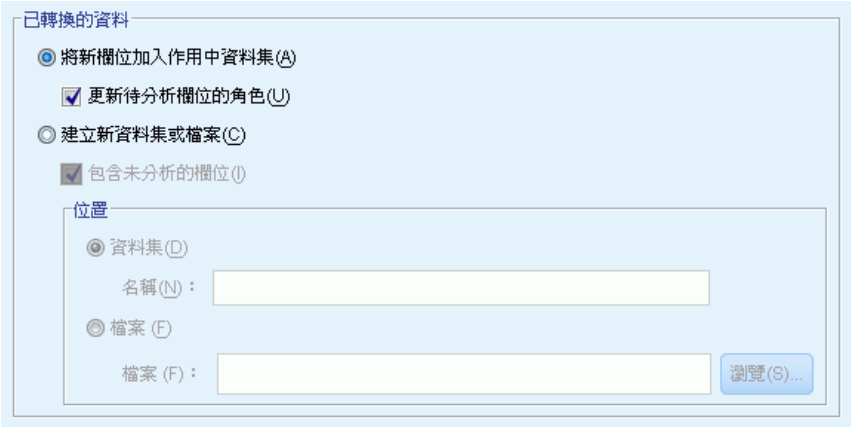
從日期與時間萃取的循環元素。指定副檔名以套用至從日期與時間萃取出的循環元素。

套用並儲存轉換

根據您使用的是「互動式資料準備」或「自動資料準備」對話方塊而定，套用與儲存轉換的設定會有些許不同。

互動式資料準備的「套用轉換」設定

圖表 4-11
互動式資料準備的「套用轉換」設定



已轉換的資料。這些設定指定儲存轉換資料的位置。

- **將新欄位加入作用中資料集。**「自動資料準備」建立的任何欄位，都會新增至作用中資料集做為新欄位。「**更新待分析欄位的角色**」會將「自動資料準備」從進一步分析中排除之任何欄位的角色設為「無」。
- **建立包含已轉換資料的新資料集或檔案。**自動資料準備建議的欄位，都會新增至新資料集或檔案。「**包含未分析的欄位**」會將「欄位」索引標籤中未指定之原始資料集的欄位新增至新資料集。這對將包含建模未用資訊的欄位（例如 ID 或地址或名稱）移轉至新資料集非常實用。

自動資料準備的「套用並儲存」設定

圖表 4-12
自動資料準備的「套用並儲存」設定

「轉換資料」群組與「互動式資料準備」相同。在「自動資料準備」中，有下列其他的選項可用：

套用轉換。在「自動資料準備」對話方塊中，取消選取此選項會停用全部其他「套用」和「儲存」控制項，同時維持選擇。

將轉換儲存為語法。這會將建議的轉換以指令語法的形式儲存到外部檔案。「互動式資料準備」對話方塊沒有此控制項，因為若您按一下「貼上」，其會將轉換貼到語法視窗做為指令語法。

將轉換儲存為 XML。這會將建議的轉換以 XML 形式儲存到外部檔案，這樣便可使用 TMS MERGE 與 PMML 模式合併，或使用 TMS IMPORT 套用至另一個資料集。「互動式資料準備」對話方塊沒有此控制項，因為若您在對話方塊上方的工具列中按一下「儲存 XML」，其會將轉換儲存為 XML。

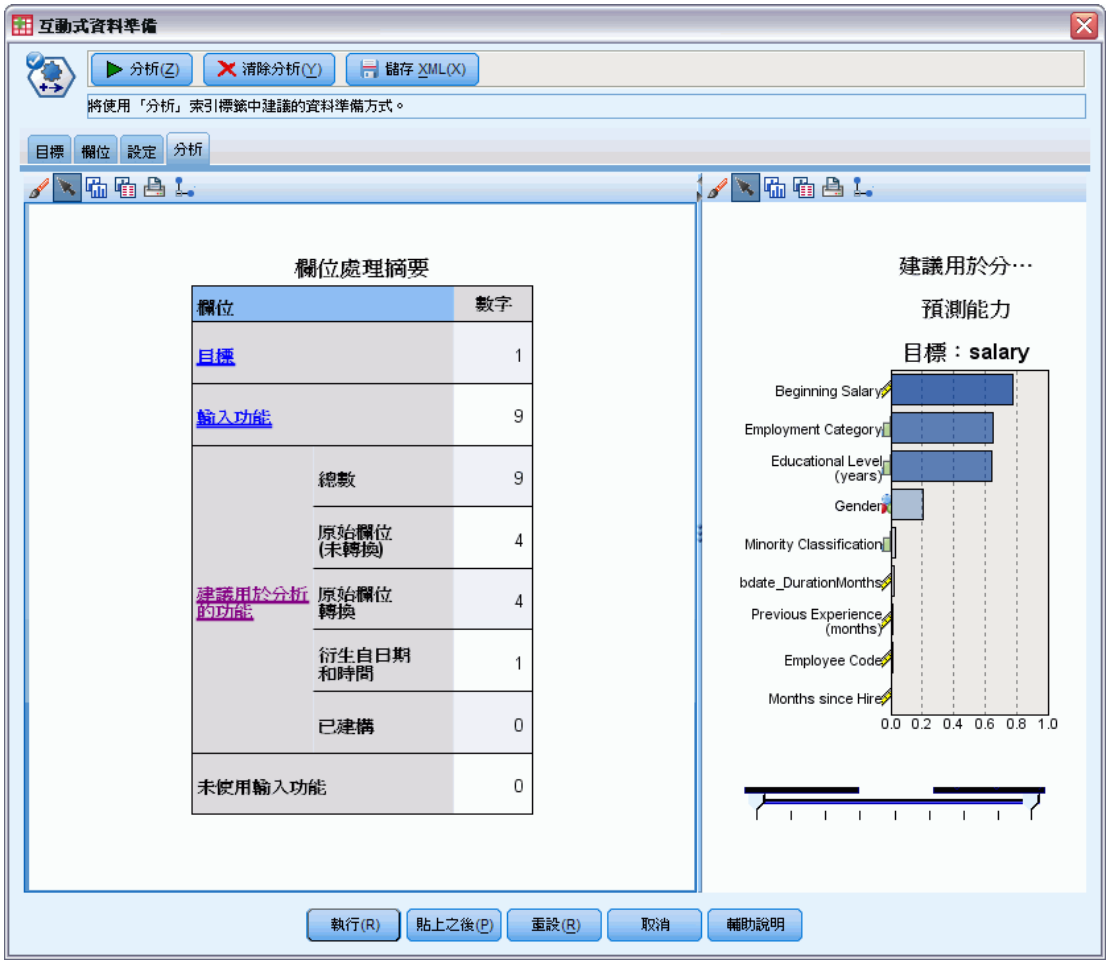
分析索引標籤

注意：「互動式資料準備」對話方塊中的「分析」索引標籤可讓您檢視建議的轉換。「自動資料準備」對話方塊則不包含此步驟。

- 當 ADP 設定（包括對「目標」、「欄位」及「設定」索引標籤的任何變更）滿足您的需求時，請按一下「**分析資料**」；演算法會將設定套用至資料輸入，並在「分析」索引標籤中顯示結果。

「分析」索引標籤包含表格和圖形輸出，這些輸出摘要說明資料的處理，並顯示關於可如何修改或改善資料以進行評分的建議。您之後可以檢視及接受或拒絕這些建議。

圖表 4-13
「自動資料準備欄位」分析索引標籤



「分析」索引標籤由兩個面板組成，主檢視位於左邊，連結或輔助檢視位於右邊。主檢視有三種：

- 欄位處理摘要（預設值）。
- 欄位。
- 動作摘要。

連結/輔助檢視有四種：

- 預測能力（預設值）。
- 欄位表格。

- 欄位詳細資料。
- 動作詳細資料。

檢視之間的連結

在主檢視中，表格內加底線的文字會控制連結檢視中的顯示。按一下文字可讓您取得特定欄位、欄位集或處理步驟的詳細資料。您最後選取的連結會以較暗的顏色顯示，這可協助您識別兩個檢視面板內容之間的關係。

重設檢視

若要重新顯示原始的「分析」建議並捨棄您對「分析」檢視所做的任何變更，請按一下主檢視面板下方的「重設」。

欄位處理摘要

圖表 4-14
欄位處理摘要

欄位處理摘要

欄位		N
目標		1
預測變數		9
建議用於分析的預測變數	總數	8
	原始欄位 (未轉換)	2
	原始欄位轉換	5
	衍生自日期和時間	1
	已建構	0
未使用預測變數		1

「欄位處理摘要」表格提供投射的整體處理影響快照，包括功能狀態的變更和建構的功能數目。

請注意，實際上不會建立任何模式，因此資料準備之前和之後都沒有整體預測能力的變更測量值或圖形；相反地，您可以顯示個別的建議預測值的預測能力圖形。

表格會顯示下列資訊：

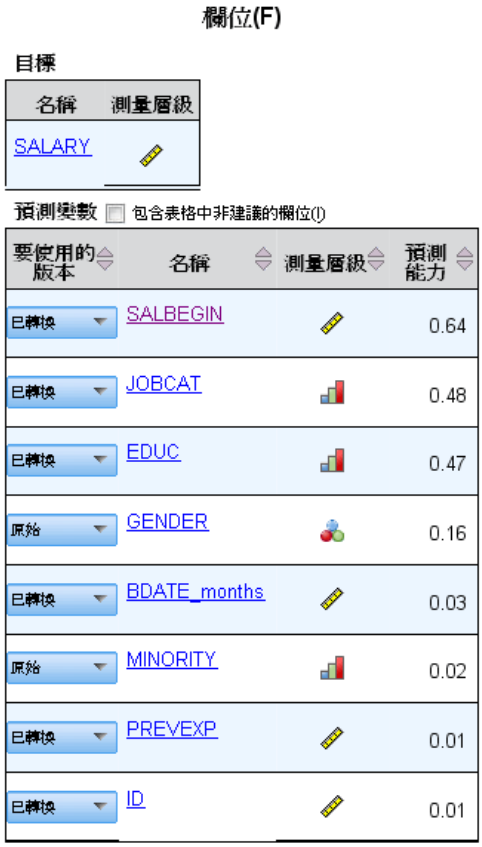
- 目標欄位數目。

- 原始（輸入）預測值的數目。
- 建議用於分析和模式建立的預測值。這包括建議的欄位總數；建議的原始、未轉換、欄位數目；建議的已轉換欄位數目（不包括任何欄位的中間版本、從日期/時間預測值衍生的欄位，以及建構的預測值）；從日期/時間欄位衍生的建議欄位數目；以及建議的已建構預測值數目。
- 輸入預測值的數目不建議以任何格式使用，無論是以其原始格式（衍生的欄位）或以建構預測值的輸入格式。

在加底線的任一「**欄位**」資訊按一下，即可在連結的檢視中顯示更多詳細資料。「**目標**」、「**輸入功能**」和「**未使用的輸入功能**」會顯示於「欄位表格」連結檢視中。 「**建議用於分析的功能**」會顯示在「預測能力」連結檢視中。

欄位

圖表 4-15
欄位



「欄位」主檢視顯示處理的欄位，以及 ADP 是否建議將它們用於下游模式中。您可以覆寫任何欄位的建議；例如，排除建構的功能或包含 ADP 建議排除的功能。若欄位已經過轉換，您可以決定要接受建議的轉換或使用原始版本。

「欄位」檢視包含兩個表格，一個代表目標，一個代表已處理或建立的預測值。

目標表格

當資料中有定義目標時，才會顯示「目標」表格。

表格包含兩行：

- **名稱。**這是目標欄位的名稱或標記；原始名稱永遠會顯示，即使欄位已經過轉換也一樣。
- **測量水準。** 這會顯示代表測量水準的圖示；將滑鼠移到圖示上方即可顯示描述資料的標記（連續、次序、名義等）。

  若目標經過轉換，則**測量水準**欄會反映最終的轉換版本。注意：您無法關閉目標的轉換功能。

預測值表格

永遠都會顯示**預測值**表格。表格的每一列代表一個欄位。根據預設值，列是以預測能力的遞減順序排序。

對於一般的功能，原始名稱永遠會做為列名稱。原始和衍生版本的日期/時間欄位會顯示於表格中（以個別的列顯示）；表格也會包含建構的預測值。

請注意，表格中顯示的已轉換版本欄位永遠代表最終的版本。

依照預設值，只有建議的欄位會顯示在「預測值」表格。若要顯示其餘的欄位，請選取表格上方的「**在表格中包含非建議的欄位**」方塊；接著就會在表格下方顯示這些欄位。

表格包含下列行：

- **要使用的版本。**這會顯示下拉式清單，此下拉式清單控制欄位是否用於下游，以及是否使用建議的轉換。依照預設值，下拉式清單會反映建議。

  對於已轉換的一般預測值，下拉式清單有三個選項：「**轉換**」、「**原始**」和「**不使用**」。

  對於未轉換的一般預測值，選項為：「**原始**」和「**不使用**」。

  對於衍生的日期/時間欄位和建構的預測值，選項為：「**轉換**」和「**不使用**」。

  對於原始日期欄位，下拉式清單是停用的，並且設為「**不使用**」。

  注意：對於含有原始和轉換版本的預測值，變更**原始**和**轉換**版本會自動更新那些功能的**測量水準**和**預測能力**設定。
- **名稱。**每個欄位名稱都是一個連結。在名稱上按一下可以在連結的檢視中顯示欄位的相關資訊。
- **測量水準。** 這會顯示代表資料類型的圖示；將滑鼠移到圖示上方即可顯示描述資料的標記（連續、次序、名義等）。
- **預測能力。**只有 ADP 建議的欄位會顯示預測能力。若未定義任何目標，則不會顯示此行。預測能力範圍介於 0 到 1，較大的數值代表「較佳」的預測值。一般來說，預測能力對於在 ADP 分析內比較預測值非常實用，但不應在分析中比較預測能力值。

動作摘要

圖表 4-16
動作摘要

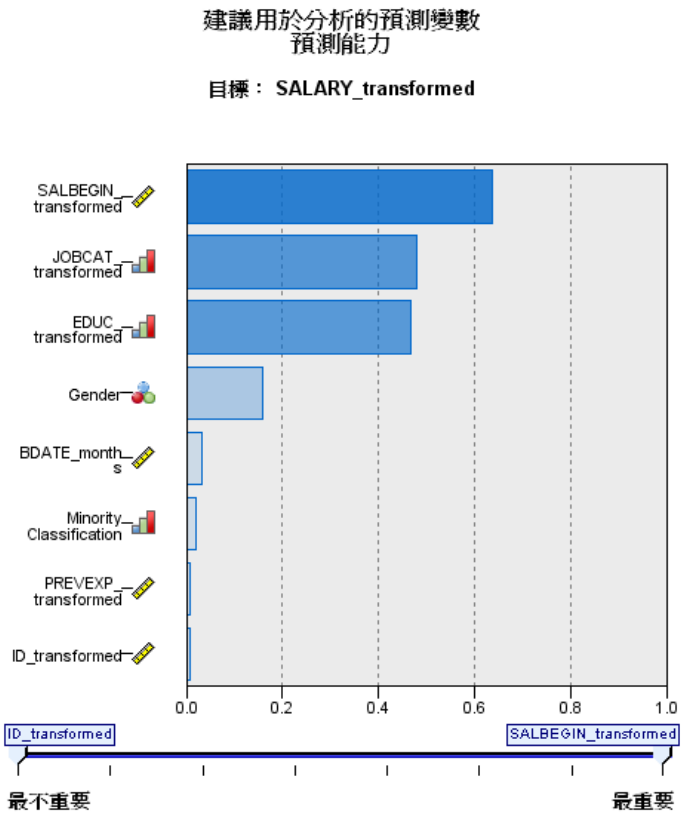
系統會針對自動資料準備所採取之各動作，轉換和/或過濾出輸入預測值；動作後留下來的欄位會用於下一個動作。之後，系統便會建議將留到最後一個步驟的欄位用於模式建立，並且過濾出轉換和建構預設值的輸入。

「動作摘要」是個簡單的表格，會列出 ADP 所採取的處理動作。按一下其中任何加底線的**動作**，便會在連結的檢視中顯示更多關於執行動作的詳細資料。

注意：只有原始和最終轉換版本的每個欄位會顯示，不會顯示分析期間使用的任何中間版本。

預測能力

圖表 4-17
預測能力



在第一次執行分析，或是選取「欄位處理摘要」主檢視的**建議用於分析的預測值**時，則會依預設顯示。此圖表顯示建議預設值的預測能力。欄位會依照預測能力排序，具有最高值的欄位會顯示於上方。

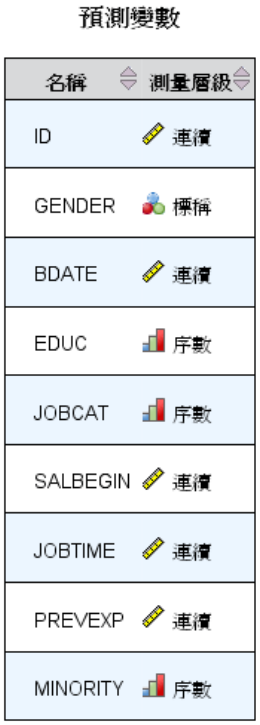
對於轉換版本的一般預測值，欄位名稱反映您在「設定」索引標籤的「欄位名稱」面板選擇的字尾；例如：_transformed。

測量水準圖示會顯示在個別的欄位名稱之後。

依據目標是連續或類別而定，系統會從線性迴歸或 Naïve Bayes 模式中計算每個建議預測值的預測能力。

欄位表格

圖表 4-18
欄位表格



當您在「欄位處理摘要」主檢視中按一下**目標**、**預測值**或**未使用的預測值**」時，就會顯示「欄位表格」檢視，其會顯示一個列出相關功能的簡單表格。

表格包含兩行：

- **名稱。** 預測值名稱。

  對於目標會使用欄位原始名稱或標記，即使目標已經過轉換也一樣。

  對於轉換版本的一般預測值，名稱會反映您在「設定」索引標籤的「欄位名稱」面板選擇的字尾；例如：_transformed。

  對於從日期與時間中衍生的欄位，會使用最終轉換版本的名稱；例如：bdate_years。

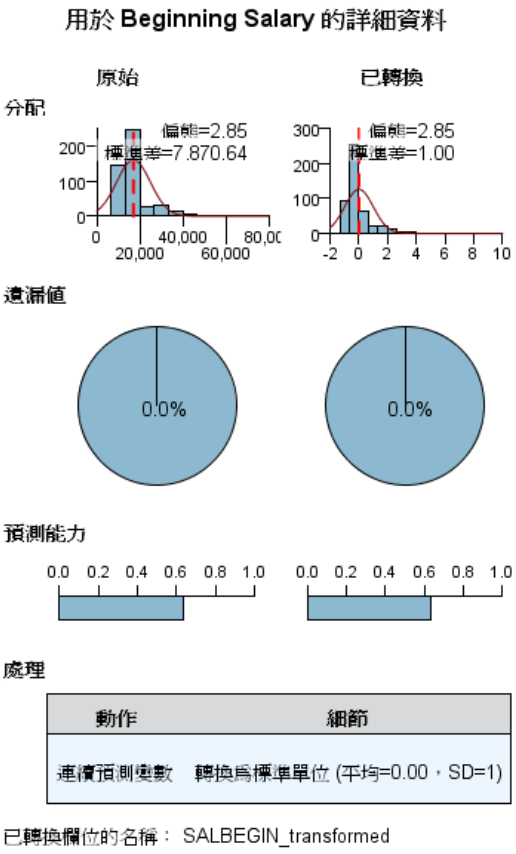
  對於建構的預測值，會使用建構預測值的名稱；例如：Predictor1.

- **測量水準。**這會顯示代表資料類型的圖示。

  對於目標，**測量水準**永遠反映轉換的版本（若目標已經過轉換）；例如，從次序（排序集合）變更為連續（範圍、尺度），反之亦然。

欄位詳細資料

圖表 4-19
欄位詳細資料



當您在「欄位」主檢視中按一下任何「名稱」時，就會顯示「欄位詳細資料」。「欄位詳細資料」檢視包含所選欄位的分配、遺漏值或預測能力圖表（如果適用）。此外，也會顯示欄位的處理記錄和轉換欄位的名稱（如果適用）。

對於每個圖表集合，會以並排的方式t顯示兩個版本，以比較套用和未套用轉換的欄位；若轉換版本的欄位不存在，則只會顯示原始版本的圖表。對於衍生的日期或時間欄位及建構的預測值，則只會顯示新預測值的圖表。

注意：若某個欄位因為有太多類別而被排除，便只會顯示處理記錄。

分配圖表

連續欄位分配會顯示為直方圖並重疊常態曲線，而垂直參考線代表平均值；類別欄位則顯示為長條圖。

直方圖標記為顯示標準差及偏態，然而，如果值的數目少於 2 ，或者原始欄位的變異數少於 10-20，則不會顯示偏態。

將滑鼠移到圖表上方，即可顯示直方圖的平均數，或是長條圖中類別記錄總數的個數及百分比。

遺漏值圖表

圓餅圖會比較套用轉換和未套用轉換的遺漏值百分比；圖表標記會顯示百分比。

若 ADP 執行了遺漏值處理，則轉換過後的圓餅圖也會包含置換值以做為標記，也就是說，會使用此值取代遺漏值。

將滑鼠移到圖表上方，會顯示遺漏值個數與記錄總數百分比。

預測能力圖表

對於建議的欄位，長條圖會顯示轉換前後的預測能力。若目標已經過轉換，則計算的預測能力會和轉換後的目標有關。

注意：若未定義目標或是在主檢視面板中按一下目標，則不會顯示預測能力圖表。

將滑鼠移到圖表上方，會顯示預測能力值。

處理記錄表格

表格會顯示轉換版本的欄位如何衍生。ADP 執行的動作會以它們執行的順序列出；不過，某些步驟的特定欄位可能會執行多個動作。

注意：未經過轉換的欄位不會顯示此表格。

表格中的資訊分為二或三行：

- **動作。** 動作的名稱。例如「連續預測值」。
- **詳細資料。**所執行處理的清單。例如，轉換為標準單位。
- **函數。**只有建構的預測值會顯示函數。函數會顯示輸入欄位的線性組合，例如 .06*age + 1.21*height。

動作詳細資料

圖表 4-20
ADP 分析 - 動作詳細資料

連續預測變數

轉換	預測變數數量	準則	
		平均值	標準差
轉換爲標準單位	5	0.00	1.00

預測變數空間建構	N
已建構的預測變數	0
由於與目標的低度關聯，已排除預測變數	1
由於這些項目經過 Binning 後成爲常數而排除預測變數	0

當您在「動作摘要」主檢視中選取任何加底線的**動作**時，就會顯示「動作詳細資料」，「動作詳細資料」連結的檢視會顯示每個執行之處理步驟的動作特定資訊和一般資訊；系統會先顯示動作專屬的詳細資料。

針對各動作，會在連結檢視上方使用說明做為標題。動作專屬的詳細資料會顯示於標題下方，並且可能包含下列詳細資料：衍生預測值的數目、欄位重新分配、目標轉換、合併或重新排序的類別以及建構或排除之預測值。

當每個動作處理完後，處理過程中使用的預測值數目可能會變更，例如將預測值排除或合併時。

注意：若關閉某個動作或未指定任何目標，則在「動作摘要」主檢視中按一下該動作時，便會在動作詳細資料處顯示錯誤訊息。

有 9 個可能的動作，但不一定每個分析都會用到。

文字欄位表格

表格會顯示下列項目的數目：

- 從分析中排除的預測值。

日期與時間預測值表格

表格會顯示下列項目的數目：

- 從日期和時間衍生的持續期間預測值。
- 日期和時間元素。
- 衍生的日期和時間預測值總計。

若已計算任何日期持續期間，則參考日期或時間會顯示為註腳。

預測值篩選表格

此表格會顯示下列從處理排除的預測值數目：

- 常數。
- 具有太多遺漏值的預測值。
- 單一類別中具有太多觀察值的預測值。
- 具有太多類別的名義欄位（集合）。
- 篩選出的預測值總數。

檢查測量水準表格

此表格會顯示重新分配的欄位數目，內容分為：

- 次序欄位（排序集合）重新分配為連續欄位。
- 連續欄位重新分配為次序欄位。
- 總數重新分配。

如果沒有連續或次序輸入欄位（目標或預測值），這就會顯示為註腳。

偏離值表格

此表格會顯示已處理的偏離值個數。

- 根據您在「設定」索引標籤的「準備輸入與目標」面板中的設定而定，可能是已發現並刪除其偏離值的連續欄位個數，或是已發現其偏離值並設為遺漏的連續欄位個數。
- 在偏離值處理之後，因為連續欄位的個數會是常數，因此將被排除。

有一個註腳會顯示偏離值分割值；如果沒有連續的輸入欄位（目標或預測值），則會顯示另一個註腳。

遺漏值表格

此表格會顯示已置換遺漏值的欄位數目，內容分為：

- 目標。如果沒有指定目標則不會顯示此列。
- 預測值。這會進一步分為名義（集合）、次序（排序集合）及連續的數目。
- 置換的遺漏值總個數。

目標表格

這個表格會顯示目標是否已轉換，顯示為：

- Box-Cox 轉換為常態。這又進一步分為顯示指定條件（平均數和標準差）的行和 Lambda 值。
- 目標類別會重新排序以提升穩定性。

類別預測值表格

此表格會顯示下列類別預測值的數目：

- 其類別經過重新排序（最低至最高）以提升穩定性。
- 其類別經過合併以最大化和目標之關聯的功能。
- 其類別經過合併以處理稀疏類別的功能。
- 因為和目標的關聯性低而排除的功能。
- 因為合併後是常數而排除的功能。

若沒有類別預測值，則會顯示註腳。

連續預測值表格

有兩個表格。第一個顯示下列其中一項轉換的數目：

- 預測值轉換為標準單位。此外，這也會顯示轉換的預測值數目、指定的平均數以及標準差。
- 對應到一般範圍的預測值。此外，這也會顯示使用最小/最大值轉換來轉換的預測值數，以及指定的最小值與最大值。
- 經過 bin 處理的預測值與經過 bin 處理的預測值數。

第二個表格會顯示預測值空間建構詳細資料，並顯示為下列預測值的數目：

- 建構的功能。
- 因為和目標的關聯性低而排除的功能。
- 因為 bin 處理後是常數而排除的功能。
- 因為建構後是常數而排除的功能。

若沒有連續預測值為輸入，則會顯示註腳。

反向轉換分數

若某個目標已被 ADP 轉換，後續的模式會使用轉換後的目標分數和單位建立。為解讀和使用結果，您必須將預測值轉換回原始尺度。

圖表 4-21
反向轉換分數

若要反向轉換分數，在功能表中選擇：
轉換(T) > 準備建模用的資料 > 反向轉換分數(B)...

- 選取欄位以執行反向轉換。此欄位應包含轉換目標的模式預測值。
- 指定新欄位的字尾。這個新欄位將包含未轉換目標的原始尺度中的模式預測值。
- 指定包含 ADP 轉換的 XML 檔案位置。這應該是從「互動式資料準備」或「自動資料準備」對話方塊中儲存的檔案。

章節
5

識別特殊觀察值

「異常偵測」程序會搜尋以其集群標準的差異為基礎的異常觀察值。這個程序設計來以資料稽核為目的，在探索資料分析的步驟中，以及在任何推論資料分析前，快速偵測異常觀察值。這個演算法是為了一般異常偵測而設計；也就是異常觀察值的定義並非指定為任何特定的應用，例如在醫療保健產業中偵測異常付款模式或在金融產業中偵測洗錢，這些情況中可以完整定義一項異常狀況。

範例。 由於中風治療結果預測模型可能對異常觀察值很敏感，因此受雇建立這些模型的資料分析人員很擔心資料品質。某些離群值是真正獨特的觀測值，因此不適合用來預測，然而其他因資料輸入錯誤所造成的觀察值，在技術上是「正確的」，因此不會被驗證資料程序偵測到。「識別異常觀察值」程序可找出並報告這些離群值，讓分析人員可以決定如何處理它們。

統計量。 這個程序可建立對等組別、連續及類別變數的對等組別基準、以對等組別基準之離差為基礎的異常索引，及當觀察值被視為異常時影響最大之變數的變數影響數值。

資料考量

資料。 此程序可用在連續變數及類別變數上。每一列都代表一個不同的觀察，且每一行都代表對等組別所依據的不同變數。資料檔內有可用於標記輸出的觀察值識別變數，但其不會用於分析中。允許遺漏值。如果已經指定，將忽略加權變數。

偵測模式可套用至一個新的檢定資料檔。檢定資料的元素必須與訓練資料的元素相同。而且，視演算法設定而定，用於建立模式的遺漏值處理也許會在計分前套用至檢定資料檔。

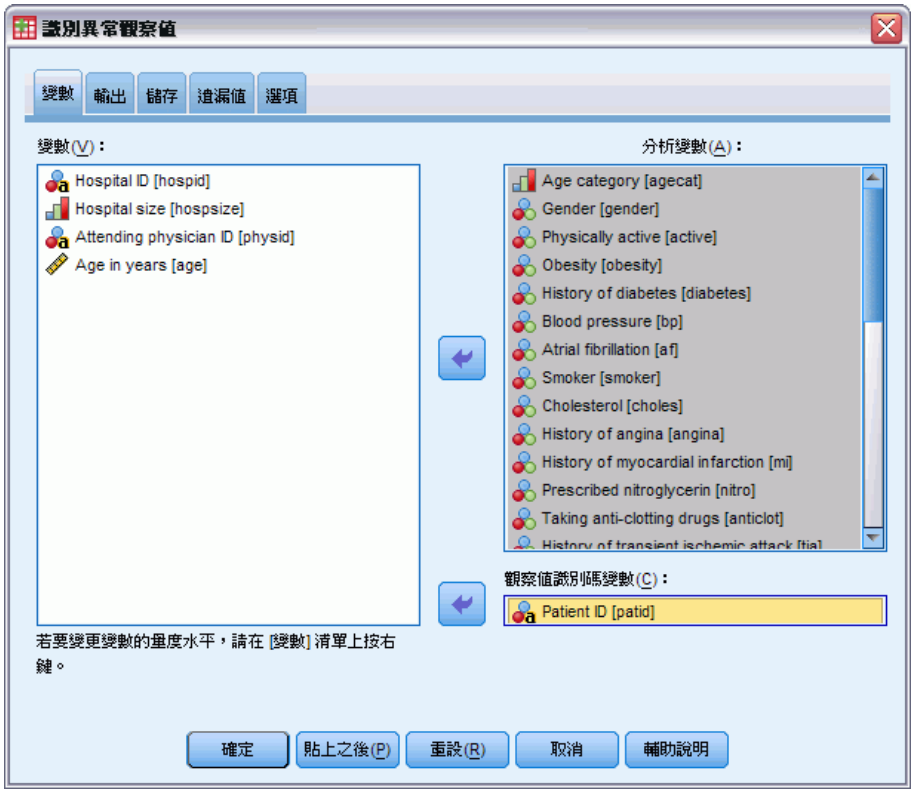
觀察值順序。 請注意解決方案可能會視觀察值順序而定。若要將順序效應降到最低，請以隨機方式排列觀察值。若要驗證某個解決方案的穩定性，您也許會想要取得幾種不同的解決方案，其觀察值皆以不同的隨機順序排列。在檔案極大的情況下，可進行多次運算，以不同的隨機順序排列一個觀察值的樣本。

假設。 演算法假設所有變數都是非常數且獨立，並假設所有觀察值在所有輸入變數中皆沒有遺漏值。每個連續變數都假設具有常態 (Gaussian) 分配，且每個類別變數都假設具有多項式分配。經驗內部檢定指出此程序很少受到獨立性假設及分配假設偏差的影響，但是要注意這些假設符合的程度。

識別異常的觀察值

▶ 從功能表選擇：
資料 > 識別特殊觀察值(I)...

圖表 5-1
「識別異常的觀察值」對話方塊，「變數」索引標籤

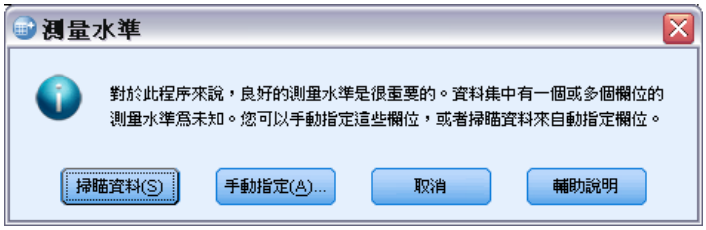


- 至少要選取一個分析變數。
- 您也可以選擇一個觀察值識別碼變數，用於標記輸出。

具有未知測量水準的欄位

若在資料集中出現一或多個未知的變數（欄位）測量水準，就會顯示「測量水準」警示。由於測量水準會影響此程序的結果計算，因此所有變數皆必須具有已定義的測量水準。

圖表 5-2
測量水準警示



- **掃描資料。** 讀取作用中資料集的資料，並且針對目前具有未知測量水準的任何欄位指派預設的測量水準。若為大型資料集，則讀取時可能需要一些時間。

- **手動指派。** 開啟對話方塊，以列出具有未知測量水準所有欄位。您可以使用此對話方塊，來指派上述欄位的測量水準。您也可以在「資料編輯程式」的「變數檢視」中指派測量水準。

由於測量水準是此程序的重要項目，因此您在所有欄位皆擁有已定義的測量水準之前，無法存取對話方塊來執行此程序。

識別異常的觀察值輸出

圖表 5-3
「識別異常的觀察值」對話方塊，「輸出」索引標籤

異常觀察值清單及它們為什麼被視為異常的原因。 此選項會產生三個表格：

- 異常觀察值索引會列出被識別為異常的觀察值，並顯示它們的對應異常索引數值。
- 異常觀察值對等 ID 清單會列出異常觀察值及其對等組別的相關資訊。
- 異常原因清單會列出每個原因的觀察值號碼、原因變數、變數影響數值、變數數值及變數的基準。

所有的表格皆以遞減的順序由異常索引排列。此外，如果「變數」索引標籤指定了觀察值識別碼變數，則會顯示觀察值的 ID。

摘要。 這個群組內的控制可產生分配摘要。

- **對等組別基準。** 這個選項顯示連續變數基準表格（如果分析中使用任何連續變數）及類別變數基準表格（如果分析中使用任何類別變數）。連續變數基準表格顯示每個對等組別中各連續變數的平均數及基準差。類別變數基準表格顯示每個對等組別中各類別變數的眾數（最普遍的類別）、次數及次數百分比。分析時會將連續變數的平均數及類別變數的眾數當成標基準值使用。
- **異常索引。** 異常索引摘要會顯示被視為異常程度最高之觀察值的異常索引敘述統計。
- **依分析變數而分的發生原因。** 對每個原因而言，此表格會將每個變數發生的次數及次數百分比顯示為原因。這個表格也報告每個變數中影響的敘述統計。如果「選項」索引標籤的最大原因數量設為 0，則這個選項無法使用。
- **觀察值已處理。** 觀察值處理摘要會顯示作用中資料集內所有觀察值的個數及個數百分比、分析中包括及不包括的觀察值，以及每個等對組別中的觀察值。

儲存識別異常的觀察值

圖表 5-4
「識別異常的觀察值」對話方塊，「儲存」索引標籤

儲存變數。 這個組別內的控制可讓您將模式變數儲存至作用中的資料集。您也可以選擇取代其名稱與將儲存的變數衝突的現有變數。

- **異常索引。** 以指定的變數名稱儲存每個觀察值的異常索引數值。

- **對等組別。** 以指定的變數根名稱儲存每個觀察值的對等組別 ID、觀察值個數及大小百分比。例如，如果已經指定根名稱「Peer」，則會產生「Peerid」、「PeerSize」，及「PeerPctSize」等變數。「Peerid」是觀察值的對等組別 ID，「PeerSize」是組別的大小，「PeerPctSize」是組別大小的百分比。
- **原因。** 以指定的根名稱儲存推理變數的組合。推理變數組合包括作為原因的變數名稱、其變數影響量數、其本身數值及基準數值。組合的數量視「選項」索引標籤所要求的原因數量而定。例如，若已經指定「Reason」根名稱，則會產生「ReasonVar_k」、「ReasonMeasure_k」、「ReasonValue_k」及「ReasonNorm_k」等變數，其中「k」為第「k」個原因。如果原因的數量設為 0，則無法使用這個選項。

匯出模式檔案。 可讓您以 XML 格式儲存模式。

識別異常觀察值的遺漏值：

圖表 5-5
「識別異常的觀察值」對話方塊，「遺漏值」索引標籤

「遺漏值」索引標籤會用於控制使用者遺漏及系統遺漏值的處理。

- **自分析排除遺漏值。**含有遺漏值的觀察值會從分析中排除。
- **在分析中包括遺漏值。** 連續變數的遺漏值會以其對應總平均數所取代，且類別變數的遺漏類別會組成群組並視為有效類別。已處理的變數稍後將用於分析中。或者，您可以要求建立代表每個觀察值遺漏變數比例的額外變數，並在分析中使用那個變數。

識別異常的觀察值選項

圖表 5-6
「識別異常的觀察值」對話方塊，「選項」索引標籤

識別異常觀察值的條件。 這些選擇會決定異常清單將包括多少觀察值。

- **最高異常索引數值的觀察值百分比。** 請指定一個小於或等於 100 的正數。
- **最高異常索引數值的觀察值固定數量。** 請指定一個小於或等於作用中資料集內用於分析之觀察值總數的正整數。
- **只識別其異常索引值符合或超過最低值的觀察值。** 指定一個非負數的數字。如果觀察值的異常索引數值大於或等於指定的分割點，則這個觀察值會被視為異常。這個選項會與「**觀察值百分比**」及「**觀察值固定數量**」選項一起使用。例如，若您指定固定數量為 50 個觀察值及分割值 2，則異常清單將包括至少 50 個觀察值，其中每個觀察值的異常索引數值都大於或等於 2。

對等組別的個數。 這個程序將在最小及最大指定值間搜尋對等組別的最佳個數。這項數值必須為正整數，而且最小值不得超過最大值。指定數值相等時，這個程序會假設對等組別的固定數量。

注意：視您資料中差異的數量而定，可能在某些情況下，資料可支援的對等組別數量小於指定的最小數量。在這種情況下，這個程序可能會建立數量較少的對等組別。

最大原因數。 一個原因會包括變數影響量數、原因的變數名稱、變數的數值及對應對等組別的數值。請指定一個非負數的整數；如果這個數值等於或大於用於分析中之已處理變數的數量，則會顯示所有變數。

DETECTANOMALY 指令的其他功能

指令語法語言也可以讓您：

- 不需明確指定所有分析變數，於分析時略過作用中資料集的幾個變數(使用「EXCEPT」次指令)。
- 指定調整以平衡連續及類別變數的影響（使用「CRITERIA」次指令中的「MLWEIGHT」關鍵字)。

如需完整的語法資訊，請參閱《指令語法參考手冊》。

章節
6

最適 Binning

「最適 Binning」程序把各個變數的值分散到 Bin 中，將一或多個尺度變數離散化（因此稱做 **Binning 輸入變數**）。對於「監督」 Binning 程序的類別導引變數而言，Bin 的構成最為適當。接著可使用 Bin 取代原始資料值做進一步分析。

範例。 將變數取得不同值的數目減少，有一些用途，包括：

- 其他程序的資料需求。離散化變數可視為類別變數，以供需要類別變數的程序使用。例如，交叉表程序需要所有的變數均為類別變數。
- 資料隱私性。報告 Bin 值而非實際值，可保護資料來源的隱私性。「最適 Binning」程序可引導選擇 Bin。
- 效能加速。有些程序在使用的不同值數目減少時更有效率。例如，使用離散化變數可改善「多項式 Logistic 迴歸」的速度。
- 找出完成或似是完成的資料分組。

最適 Binning 對 Visual Binning。「Visual Binning」對話方塊提供數個自動方法可建立 Bin，而不使用導引變數。這些「未受監督」的規則對於產生敘述統計很有用，例如次數表，但當您最終目標是產生預測模式時，「最適 Binning」是較佳的選擇。

輸出。 此程序會為 Bin 產生分割點表，並為每個 Binning 輸入變數產生敘述統計。此外，您可以將新變數儲存至包含 Binning 輸入變數其之 Bin 值的作用中資料集，並將 Binning 規則儲存為指令語法，用來離散化新資料。

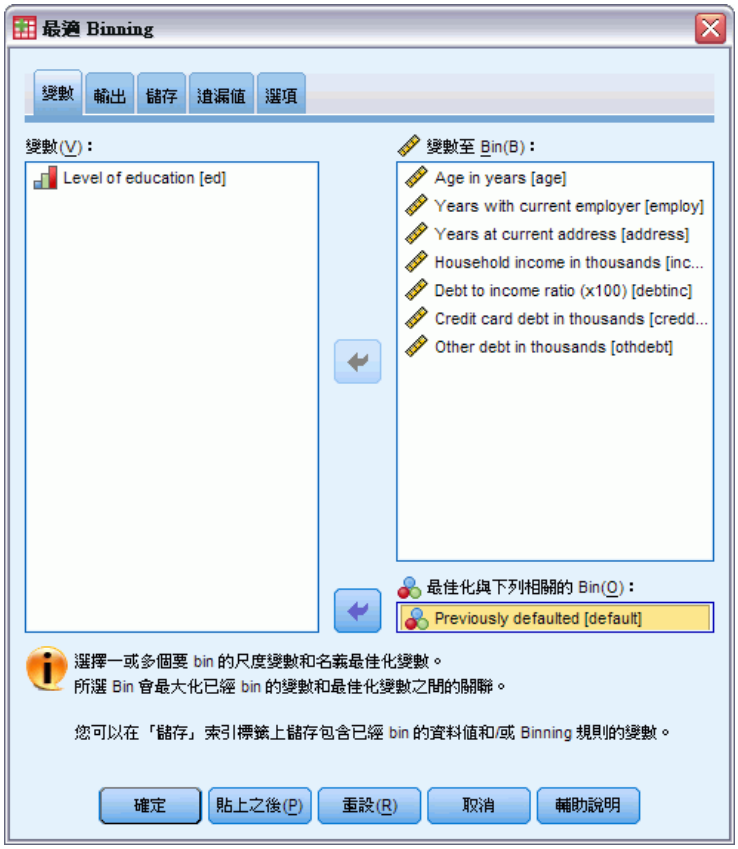
資料。 此程序預期 Binning 輸入變數為尺度、數值變數。導引變數應為類別變數，且可以是字串或數值。

若要取得最適 Binning

從功能表選擇：

轉換 > 最適 Binning...

圖表 6-1
「最適 Binning」對話方塊，「變數」索引標籤

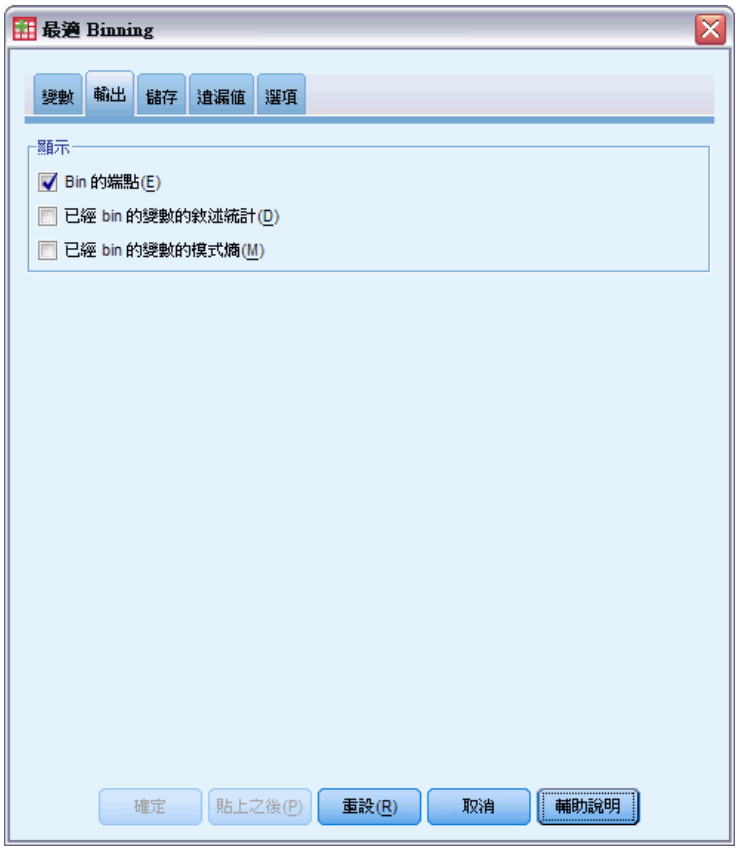


- 選取一或多個 Binning 輸入變數。
- 選取導引變數。

根據預設值，不會產生包含 Bin 資料值的變數。使用「儲存」索引標籤儲存這些變數。

最適 Binning 輸出

圖表 6-2
「最適 Binning」對話方塊，「輸出」索引標籤



「輸出」索引標籤可控制結果的顯示。

- **Bin 的端點。** 顯示每個 Binning 輸入變數的端點集。
- **已經過 Bin 處理變數的敘述統計。** 對於每個 Binning 輸入變數，此選項會顯示具有效值的觀察值數目、具遺漏值的觀察值數目、不同有效值的數目，和最小值與最大值。對於導引變數，此選項會顯示每個相關 Binning 輸入變數的類別分配。
- **已經過 Bin 處理變數的模式熵。** 對於每個 Binning 輸入變數，此選項會顯示與導引變數有關的變數預測準確性測量。

最適 Binning 儲存

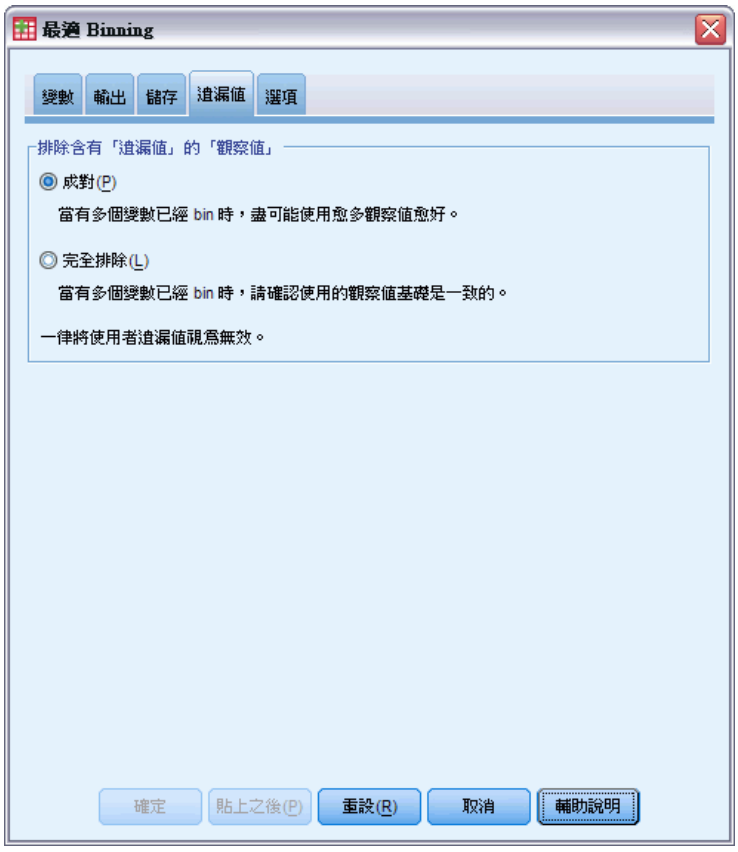
圖表 6-3
「最適 Binning」對話方塊，「儲存」索引標籤

將變數儲存至作用中資料集。 在進一步分析時，可使用包含 Bin 資料值的變數來取代原始變數。

將 Binning 規則儲存為語法。 產生可用於 Bin 其他資料集的指令語法。記錄規則是根據由 Binning 演算法決定的分割點。

最適 Binning 遺漏值

圖表 6-4
「最適 Binning」對話方塊，「遺漏值」索引標籤



「遺漏值」索引標籤會指定使用完全排除或成對刪除處理遺漏值。使用者遺漏值一律視為無效。將原始變數值記錄至新變數時，會將使用者遺漏值轉換為系統遺漏值。

- **成對。** 此選項會在每個導引變數和 Binning 輸入變數配對上作業。程序將利用所有包含導引變數和 Binning 輸入變數上非遺漏值的觀察值。
- **完全排除** 此選項會在於「變數」索引標籤上指定的所有變數間操作。如果觀察值有任何變數遺漏，則會排除整個觀察值。

最適 Binning 選項

圖表 6-5
「最適 Binning」對話方塊，「選項」索引標籤

預先處理。 若「預先 Binning」的 Binning 輸入變數有許多不同值，則該變數可以改善處理時間，而不必大幅犧牲最後 Bin 的品質。Bin 的最大數為建立 Bin 數目的上限。因此，如果您指定 1000 為最大值，但 Binning 輸入變數的不同值少於 1000，則為 Binning 輸入變數建立的預先處理 Bin 將等於 Binning 輸入變數的不同值數目。

稀疏集合的 Bin。 該程序偶爾會產生觀察值非常少的 Bin。以下策略會刪除這些虛擬分割點：

- 對於指定的變數，假設演算法發現有 n 個 final 分割點，因此 n_{final}+1 Bin。對於 Bin i = 2, ..., n_{final}（第二低值 Bin 到第二高值 Bin），計算

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}),sizeof(b_{i+1}))}$$

其中 sizeof(b) 是 Bin 中的觀察值數目。

- 當此值低於指定的合併門檻時，b_i 則視為稀疏集合，並與 b_{i-1} 或 b_{i+1} 合併，無論哪一個都有較低的類別資訊模式熵。

此程序會產生單一通道穿過 Bin。

Bin 端點。 此選項會指定如何定義區間的下限。由於程序會自動決定分割點值，這主要是偏好的問題。

第一（最低）/ 最後（最高）Bin。 這些選項指定如何定義每個 Binning 輸入變數的最小和最大分割點。一般來說，此程序會假設 Binning 輸入變數可以在實數線上取得任何值，但如果您有一些理論上或實務上的理由要限制範圍，您可以使用最低 / 最高值來限制範圍。

OPTIMAL BINNING 指令和其他功能

指令語法語言讓您也可以：

- 透過相同次數方法執行未受監督的 Binning（使用 CRITERIA 次指令）。

如需完整的語法資訊，請參閱《指令語法參考手冊》。

部 II:
範例

章節
7

驗證資料

「驗證資料」程序可找出可疑和無效的觀察值、變數和資料值。

驗證醫學資料庫

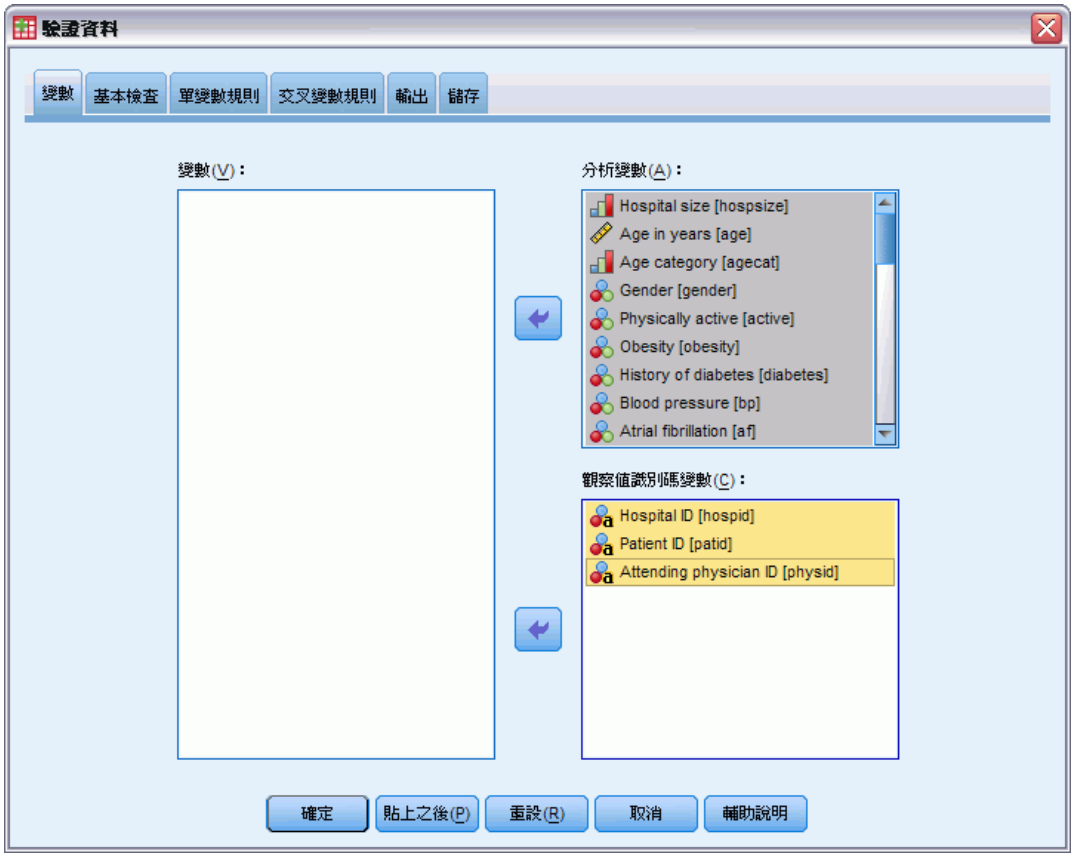
由某家醫療團體僱用的分析人員必須維持系統中資訊的品質。這個程序需要檢查數值和變數，並準備要向資料輸入小組主管呈報的報告。

資料庫的最新狀況收錄在 stroke_invalid.sav 中。 使用「驗證資料」程序取得製作報告所需的資料。進行這些分析的語法存放在 validatedata_stroke.sps 內。

執行基本檢查

- 若要執行「驗證資料」分析，從功能表中選擇：
  資料 > 驗證(V) > 驗證資料(V)...

圖表 7-1
「驗證資料」對話方塊，「變數」索引標籤



- 選擇「醫院大小」和「年齡（年為單位）」到「已記錄 6 個月的巴氏指數」作為分析變數。
- 選擇「醫院 ID」、「病患 ID」和「主治醫師 ID」作為觀察值識別碼變數。
- 按一下「**基本檢查**」索引標籤。

圖表 7-2
「驗證資料」對話方塊，「基本檢查」索引標籤

預設設定為您要執行的設定。

► 按一下「確定」。

警告

圖表 7-3
警告

所要求的輸出有一些或全部都未顯示出來，因爲所有觀察值、變數或資料值都通過所要求的檢查。

分析變數會通過基本檢查，而且沒有空觀察值，所以會顯示警告解釋沒有與這些檢查對應之輸出的原因。

不完整識別碼

圖表 7-4
不完整觀察值識別碼

	識別符		
觀察值	Hospital ID	Patient ID	Attending physician ID
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

當在觀察值識別變數中有遺漏值時，便無法適當的識別觀察值。在此資料檔案中，觀察值 288 遺漏了病患 ID，而觀察值 573 和 774 遺漏了醫院 ID。

重複的識別碼

圖表 7-5
重複觀察值識別碼（顯示前 11 個）

			識別符		
重複的識別符群組	重複次數	具有重複識別符的觀察值	Hospital ID	Patient ID	Attending physician ID
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329

觀察值應由專屬的識別碼變數值組合加以識別。在此顯示重複識別碼表中的前 11 個項目。這些重複是因為有些病患有多個事件，他們在每個事件中被輸入成不同的觀察值。因為此資訊會收錄在單一列中，所以應該清除這些觀察值。

複製和使用其他檔案中的規則

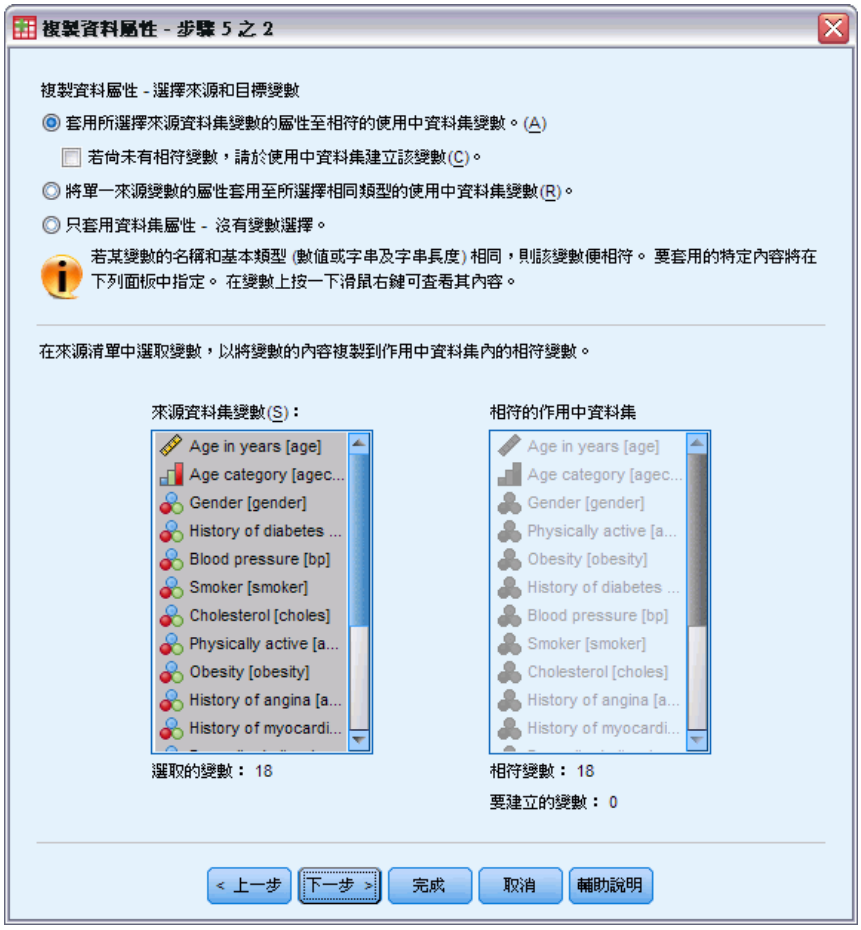
分析人員注意到此資料檔案中的變數與另一個專案中的變數類似。為該專案所定義的驗證規則會儲存為相關資料檔案的性質，藉由複製檔案的資料性質，可將規則套用到此資料檔上。

▶ 若要複製其他檔案的規則，從功能表中選擇：
資料 > 複製資料性質(C)...

圖表 7-6
複製資料性質，步驟 1（歡迎）

- 選擇從外部 IBM® SPSS® Statistics 資料檔 patient_los.sav 複製性質。
- 按一下「**下一步**」。

圖表 7-7
複製資料性質，步驟 2（選擇變數）



這些是您要將其性質從 patient_los.sav 複製到 stroke_invalid.sav 中對應變數上的變數。

- 按一下「下一步」。

圖表 7-8
複製資料性質，步驟 3（選擇變數性質）

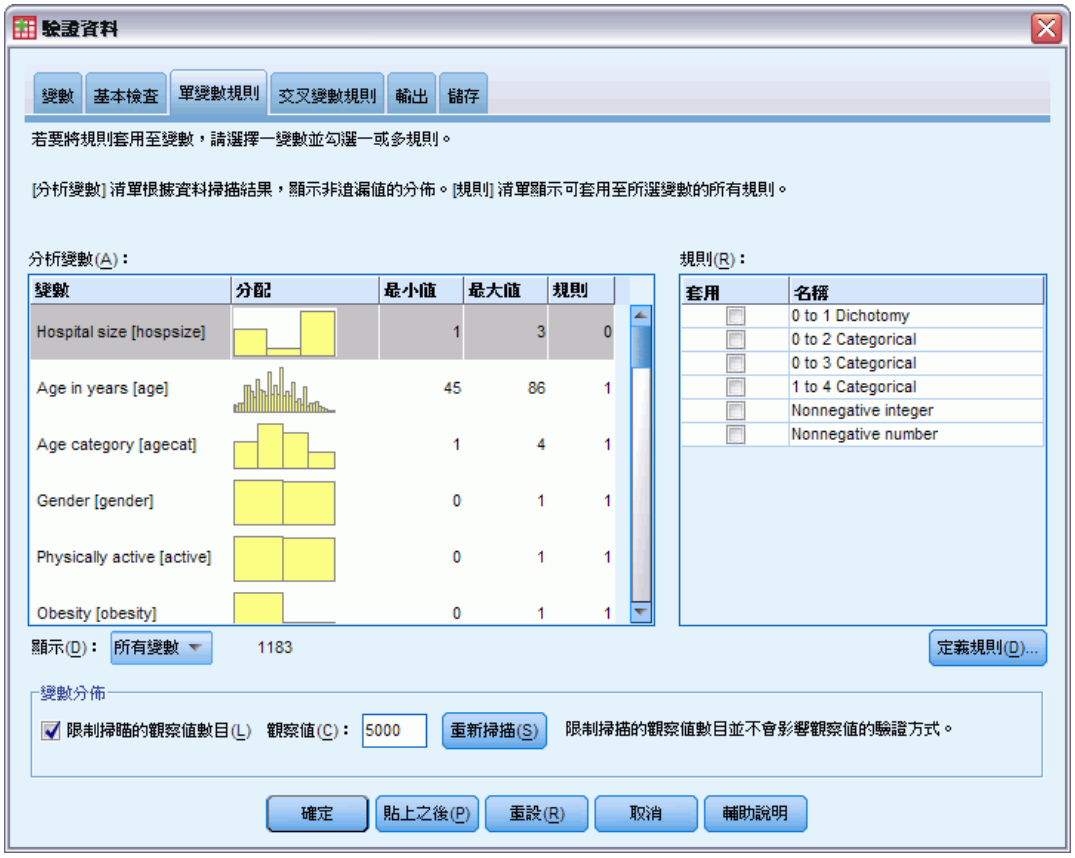
- 取消選擇所有性質，「自訂屬性」除外。
- 按一下「下一步」。

圖表 7-9
複製資料性質，步驟 4（選擇資料集性質）

- 選擇「自訂屬性」。
- 按一下「完成」。

現在您已經準備好重複使用驗證規則。

圖表 7-10
「驗證資料」對話方塊，「單一變數規則」索引標籤

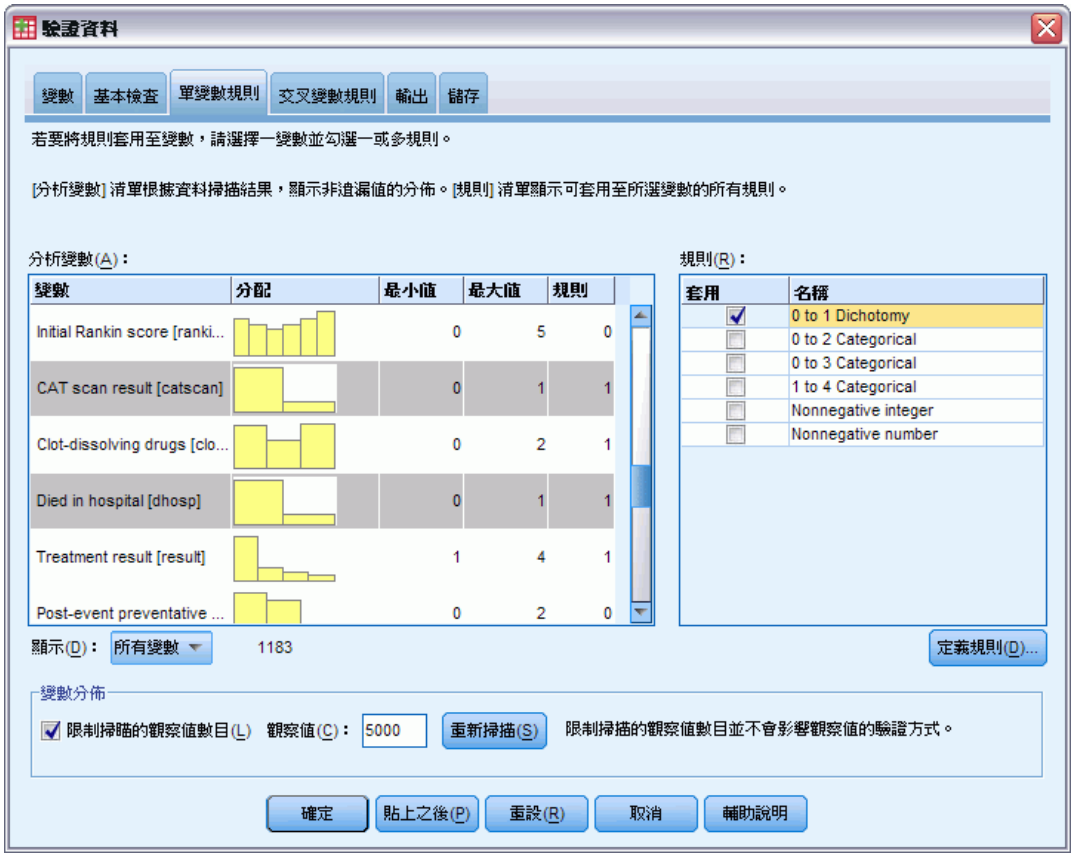


- 若要使用複製的規則來驗證 stroke_invalid.sav 資料，按一下「叫回對話」工具列按鈕，並選擇「**驗證資料**」。
- 按一下「**單一變數規則**」索引標籤。

「分析變數」清單顯示在「變數」索引標籤上選擇的變數、其分配狀態的摘要資訊，以及每個變數所附加的規則數。其性質是從 patient_los.sav 複製的變數會附有一些規則。

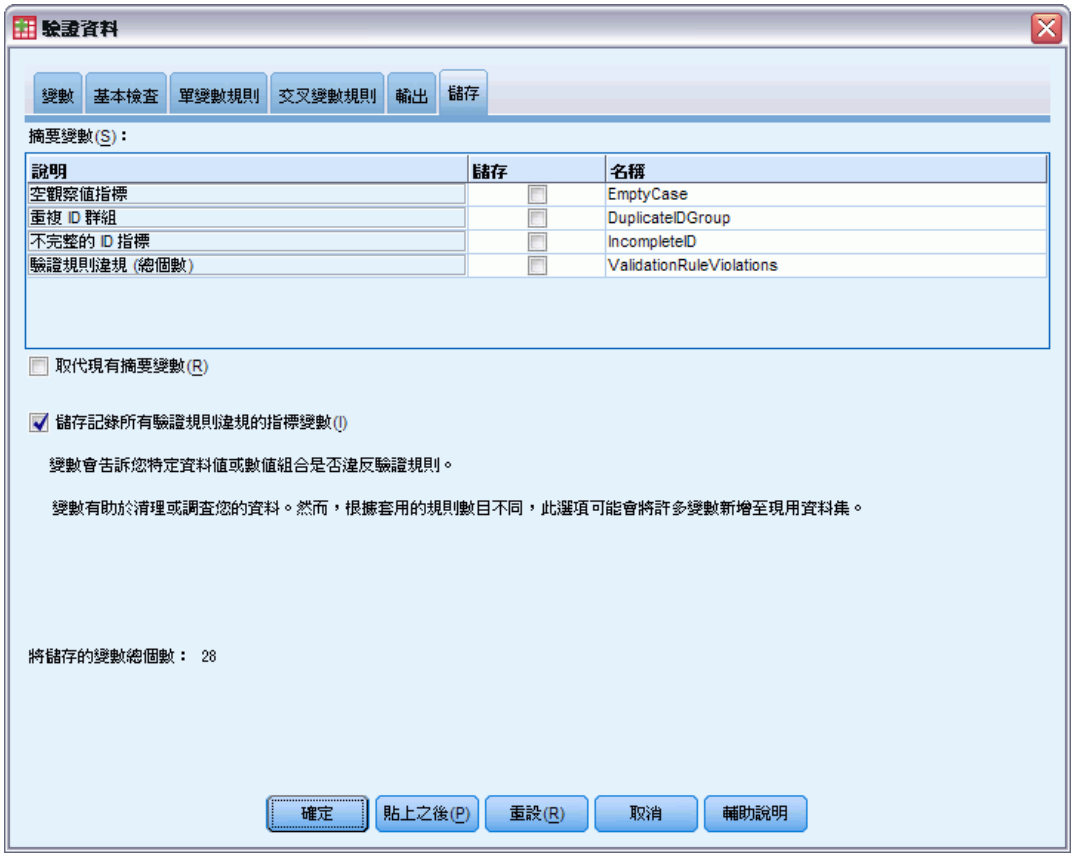
「規則」清單會顯示資料檔案中可用的單一變數驗證規則。這些規則全部複製自 patient_los.sav。請注意，這些規則有些適用於在其他資料檔案中無確實對應部分的變數。

圖表 7-11
「驗證資料」對話方塊，「單一變數規則」索引標籤



- 選擇「心房纖維顫動」、「暫時性腦缺血病史」、「CAT 掃描結果」和「院內死亡」，再套用「**0 到 1 二分集**」規則。
- 將「**0 到 3 類別量數**」套用到「事件後康復」。
- 將「**0 到 2 類別量數**」套用到「事件後預測性診療」。
- 將「**非負數整數**」套用至「復健住院日數」。
- 將「**1 到 4 類別量數**」套用到「記錄 1 個月的巴氏指數」至「記錄 6 個月的巴氏指數」。
- 按一下「**儲存**」索引標籤。

圖表 7-12
「驗證資料」對話方塊，「儲存」索引標籤



- 選擇「儲存記錄所有驗證規則違規的指標變數」。此程序可以更容易的將單一變數規則違規的觀察值和變數連接在一起。
- 按一下「確定」。

規則說明

圖表 7-13
規則說明

規則	說明
Nonnegative integer	類型:數值 網域:範圍 旗標使用者遺漏值:否 旗標系統遺漏值:是 最小:0 旗標範圍內未標記的值:否 旗標範圍內非整數值:是 $VD.SRule[5]:規則
0 to 1 Dichotomy	類型:數值 網域:清單 旗標使用者遺漏值:否 旗標系統遺漏值:是 清單:0, 1 $VD.SRule[1]:規則
1 to 4 Categorical	類型:數值 網域:清單 旗標使用者遺漏值:否 旗標系統遺漏值:是 清單:1, 2, 3, 4 $VD.SRule[4]:規則

違反了至少一次的規則會顯示出來。

規則說明表顯示違規的解釋。此功能在持續追蹤大量驗證規則上非常的有用。

變數摘要

圖表 7-14
變數摘要

	規則	違反數目
agecat	1 to 4 Categorical	1
	總和	1
gender	0 to 1 Dichotomy	1
	總和	1
angina	0 to 1 Dichotomy	1
	總和	1
time	Nonnegative integer	2
	總和	2
doa	0 to 1 Dichotomy	1
	總和	1

變數摘要表列出至少一個驗證規則違規的變數、違反的規則，以及每個規則和每個變數所發生的違規數。

觀察值報告

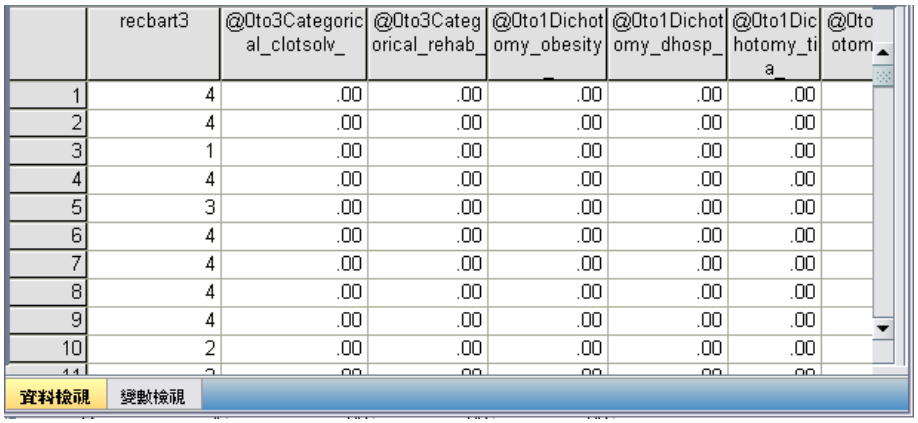
圖表 7-15
觀察值報告

	驗證規則違反	識別符		
觀察值	單一變數[a]	hospid	patid	physid
175	0 to 1 Dichotomy (1)	OZN	0333204686	883285
274	0 to 1 Dichotomy (1)	OZN	1038840465	103254
310	Nonnegative integer (1)	OZN	2090290204	883285
437	0 to 1 Dichotomy (1)	WPA	2349729006	723384
752	Nonnegative integer (1)	GFG	4993307441	828754

a. 違反規則的變數數量會出現在每個規則之後。

觀察值報告表列出至少一個驗證規則違規的觀察值（依照觀察值號碼和觀察值識別碼）、違反的規則，以及觀察值的違規數。在「資料編輯程式」中顯示無效值。

圖表 7-16
具有所儲存違規指標的「資料編輯程式」

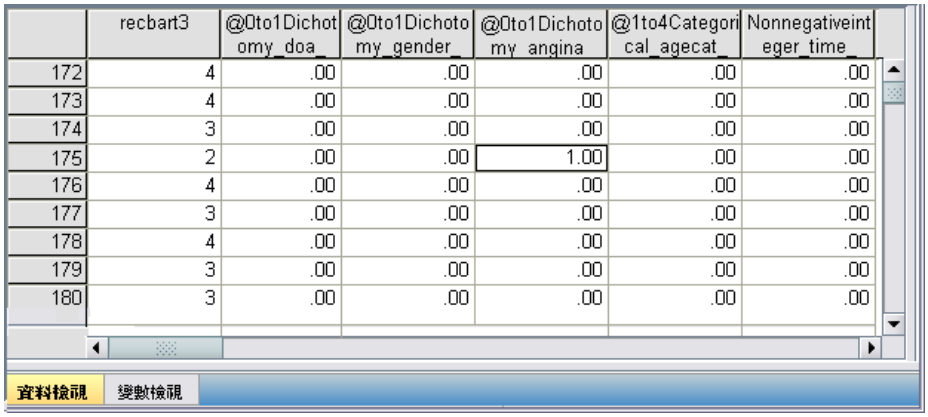


	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity_	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_	@0tootom
1	4	.00	.00	.00	.00	.00	
2	4	.00	.00	.00	.00	.00	
3	1	.00	.00	.00	.00	.00	
4	4	.00	.00	.00	.00	.00	
5	3	.00	.00	.00	.00	.00	
6	4	.00	.00	.00	.00	.00	
7	4	.00	.00	.00	.00	.00	
8	4	.00	.00	.00	.00	.00	
9	4	.00	.00	.00	.00	.00	
10	2	.00	.00	.00	.00	.00	

資料檢視 變數檢視

系統會為驗證規則的每一個套用產生不同的指標變數。因此，@0to3Categorical_clotsolv_ 是指將「0 到 3 類別量數」單一變數驗證規則套用到變數血塊溶解藥物。對於指定的觀察值，了解哪一個變數值無效的最簡單方法即是掃描指標值。指標值為 1 表示關聯的變數值無效。

圖表 7-17
具有觀察值 175 違規指標的「資料編輯程式」

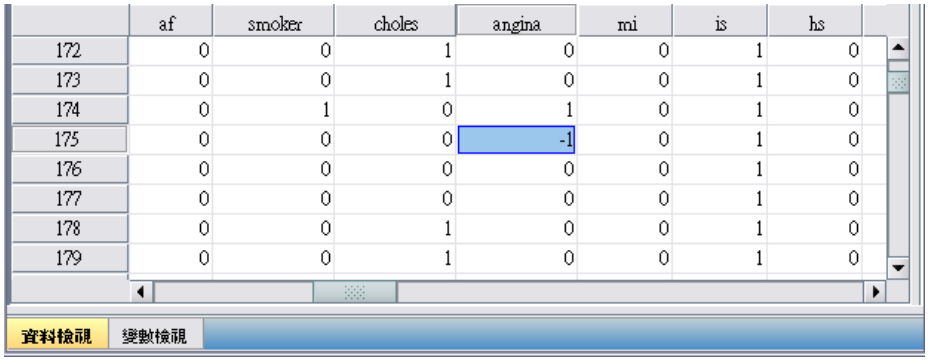


	recbart3	@0to1Dichotomy_doa_	@0to1Dichotomy_gender_	@0to1Dichotomy_angina	@1to4Categorical_agecat_	Nonnegativeinteger_time_
172	4	.00	.00	.00	.00	.00
173	4	.00	.00	.00	.00	.00
174	3	.00	.00	.00	.00	.00
175	2	.00	.00	1.00	.00	.00
176	4	.00	.00	.00	.00	.00
177	3	.00	.00	.00	.00	.00
178	4	.00	.00	.00	.00	.00
179	3	.00	.00	.00	.00	.00
180	3	.00	.00	.00	.00	.00

資料檢視 變數檢視

移至觀察值 175，也就第一個違規的觀察值。若要加快搜尋，請先看一下與變數摘要表中與變數關聯的指標。您會很容易的看到心絞痛病史有無效值。

圖表 7-18
具有「心絞痛病史」無效值的「資料編輯程式」



	af	smoker	choles	angina	mi	is	hs
172	0	0	1	0	0	1	0
173	0	0	1	0	0	1	0
174	0	1	0	1	0	1	0
175	0	0	0	-1	0	1	0
176	0	0	0	0	0	1	0
177	0	0	0	0	0	1	0
178	0	0	1	0	0	1	0
179	0	0	1	0	0	1	0

資料檢視 變數檢視

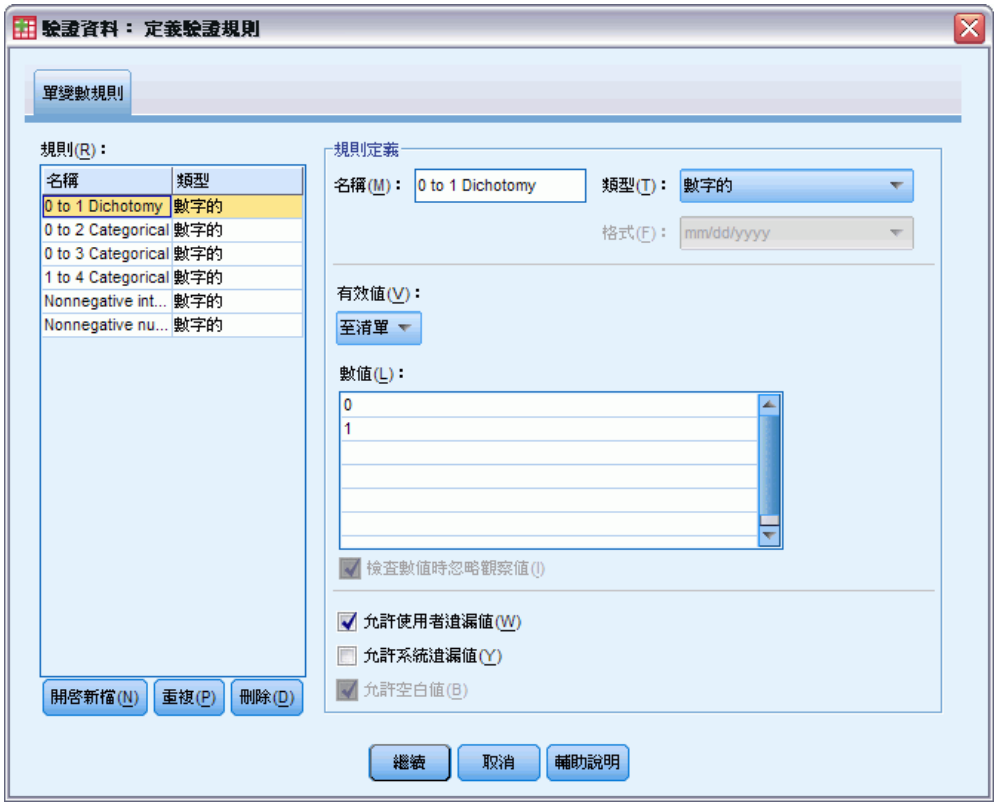
心絞痛病史的值為 -1。儘管此值是資料檔案中治療的有效遺漏值和結果變數，在此該值無效，因為病患的病史值目前沒有定義使用者遺漏值。

定義您自己的規則

複製自 patient_los.sav 的驗證規則已經很有用了，但您必須定義更多的規則來完成工作。此外，有時會不心將到院前死亡的病患標記為院內死亡。單一變數驗證規則無法滿足此狀況，所以您必須定義交叉變數規則來處理此狀況。

- 按一下「叫回對話」工具列按鈕，並選擇「**驗證資料**」。
- 按一下「**單一變數規則**」索引標籤。(您必須為測量 Rankin 分數的變數醫院大小與未記錄巴氏指數對應的變數，定義規則)。
- 按一下「**定義規則**」。

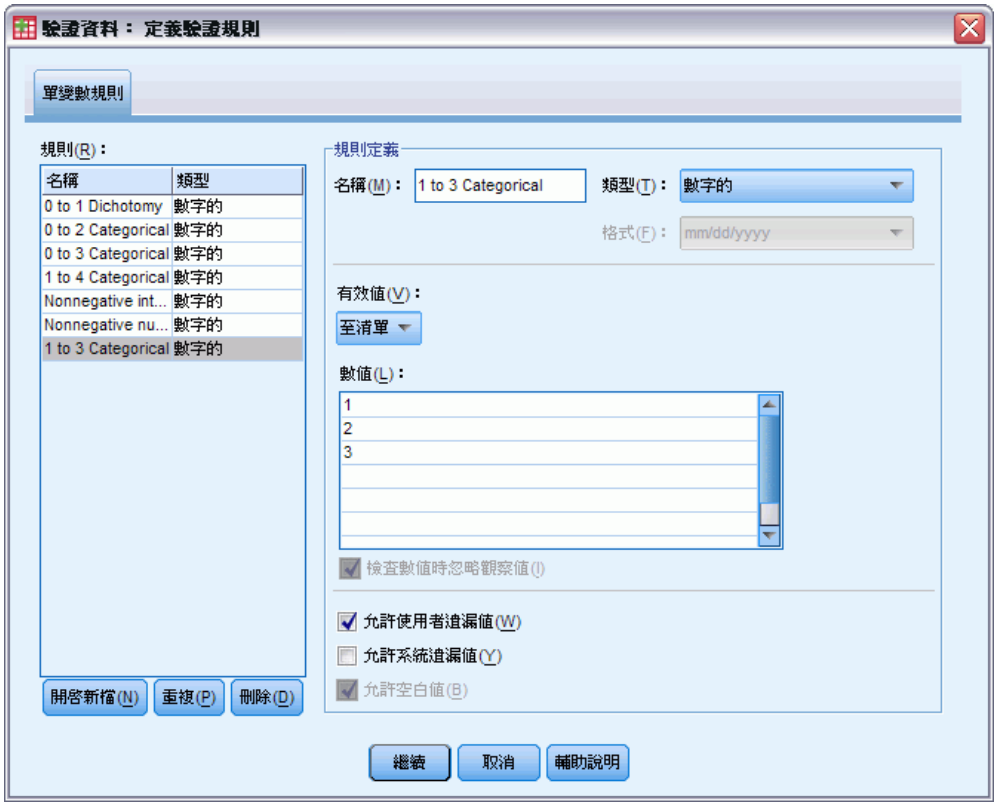
圖表 7-19
「定義驗證規則」對話方塊，「單一變數規則」索引標籤



目前定義的規則會與在「規則」清單選擇的 **0 到 1 二分集**，以及在「規則定義」組別中顯示的規則性質一起顯示。

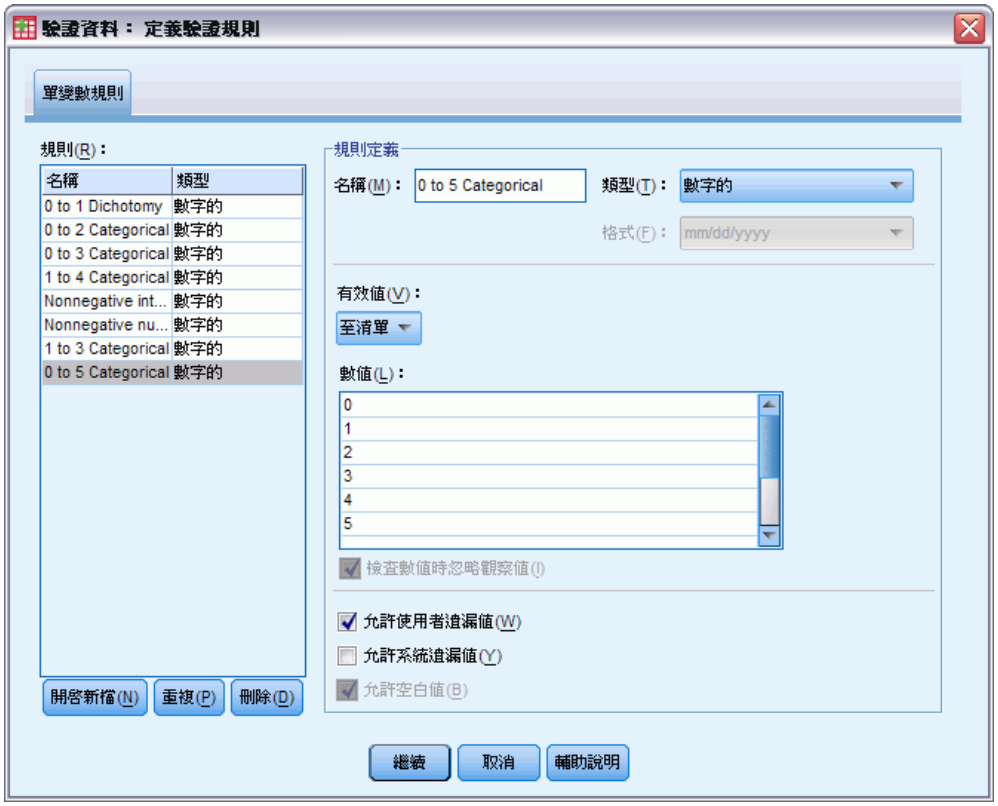
- 若要定義規則，按一下「**新增**」。

圖表 7-20
「定義驗證規則」對話方塊，「單一變數規則」索引標籤（已定義 1 到 3 個類別）



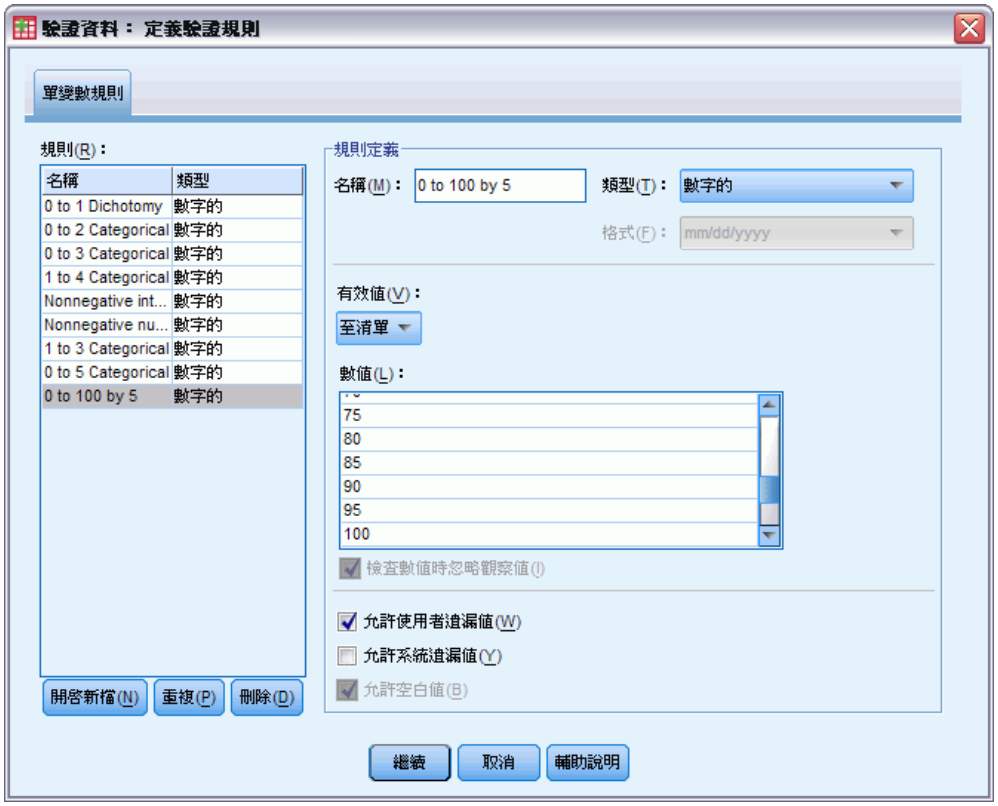
- 輸入 1 到 3 類別量數作為規則名稱。
- 對於「有效數值」，選擇「在清單中」。
- 輸入 1、2 和3 作為數值。
- 取消選擇「允許系統遺漏值」。
- 若要定義 Rankin 分數的規則，按一下「新增」。

圖表 7-21
「定義驗證規則」對話方塊，「單一變數規則」索引標籤（已定義 0 到 5 個類別）



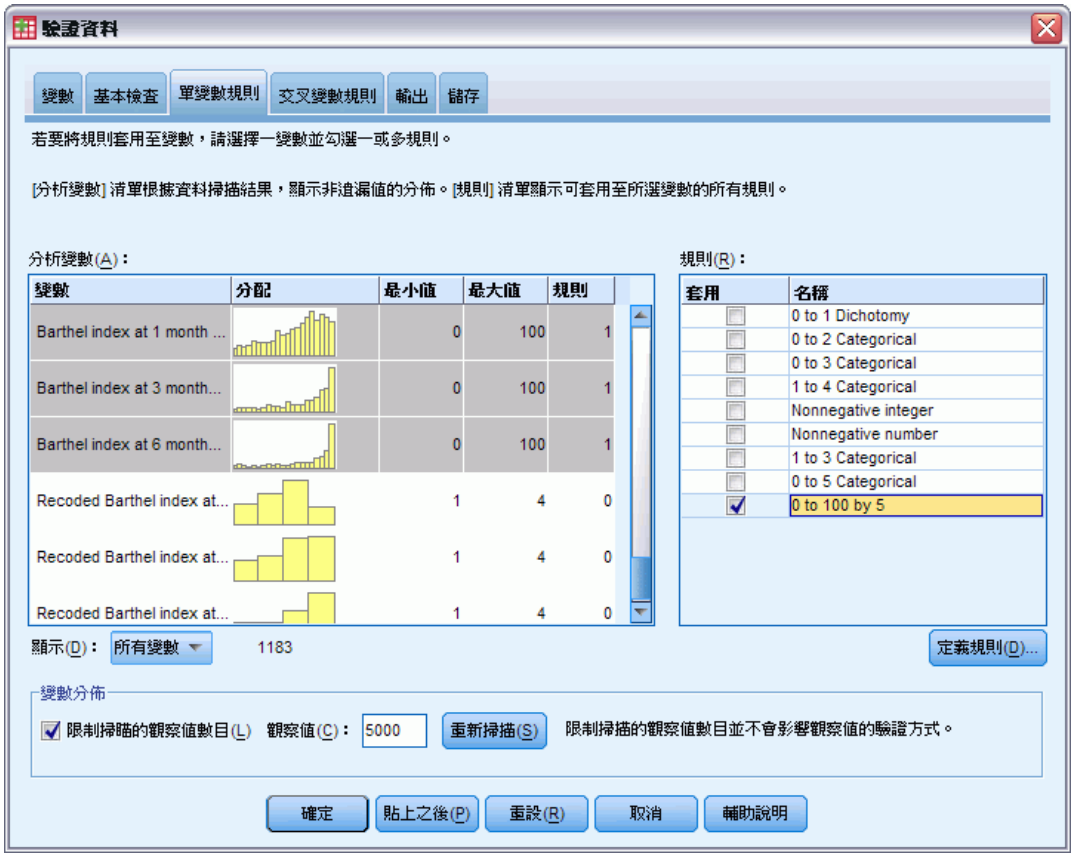
- 輸入 0 到 5 類別量數作為規則名稱。
- 對於「有效數值」，選擇「在清單中」。
- 輸入 0、1、2、3、4 和 5 作為數值。
- 取消選擇「允許系統遺漏值」。
- 若要定義巴氏指數的規則，按一下「新增」。

圖表 7-22
「定義驗證規則」對話方塊，「單一變數規則」索引標籤（已定義 0 到 100 類別，每 5 個定義為一類別）



- 輸入 0 到 100：每隔 5 作為規則名稱。
- 對於「有效數值」，選擇「在清單中」。
- 輸入 0、5. . . 和 100 作為數值。
- 取消選擇「允許系統遺漏值」。
- 按一下「繼續」。

圖表 7-23
「驗證資料」對話方塊，「單一變數規則」索引標籤（已定義 0 到 100 類別，每 5 個定義為一類別）



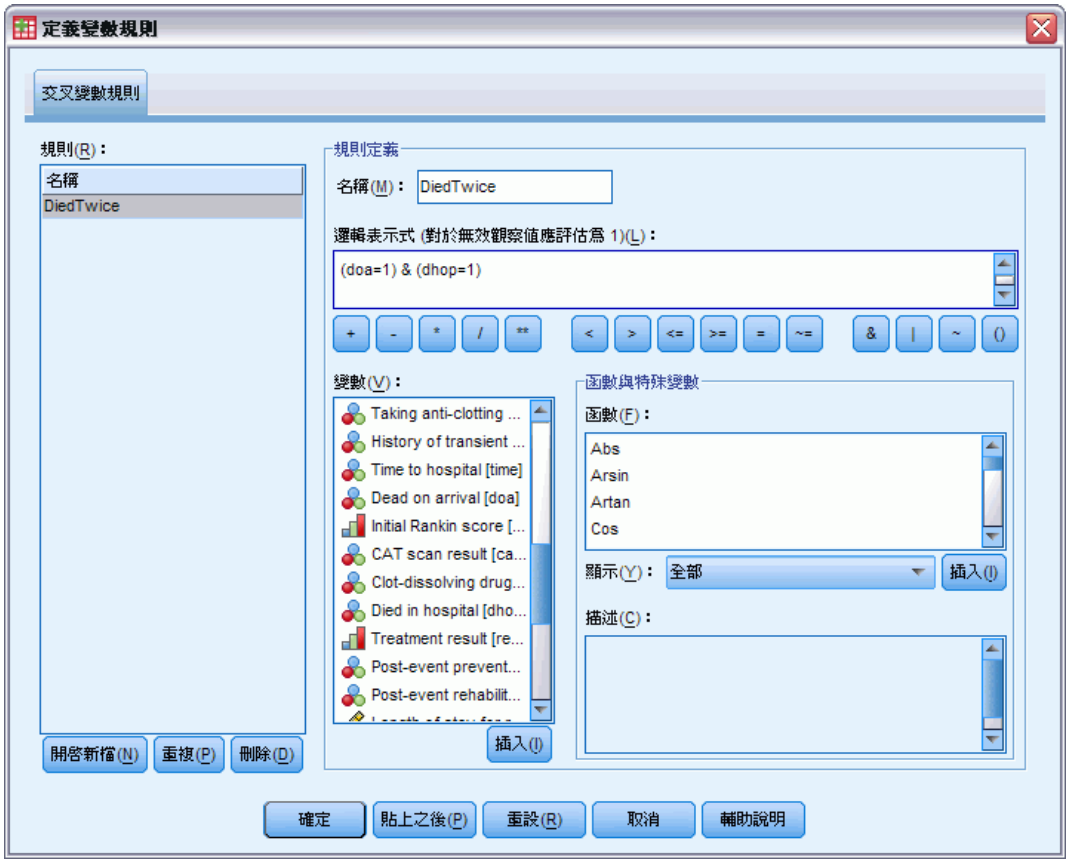
現在您必須套用已定義的規則，來分析變數。

- 將「**1 到 3 類別量數**」套用到「醫院大小」。
- 將「**0 到 5 類別量數**」套用到「開始 Rankin 分數」，和「1 個月的 Rankin 分數」到「6 個月的 Rankin 分數」。
- 將「**0 到 100：每隔 5**」套用到「1 個月的巴氏指數」到「6 個月巴氏指數」。
- 按一下「**交叉變數規則**」索引標籤。

目前沒有已定義的規則。

- 按一下「**定義規則**」。

圖表 7-24
「定義驗證規則」對話方塊，「交叉變數規則」索引標籤



當沒有規則時，會自動建立新的預留位置規則。

- 輸入 DiedTwice 作為規則名稱。
- 輸入 (doa=1) & (dhosp=1) 作為邏輯運算式。如果病患同時記錄為到院前死亡和院內死亡，則會傳回數值 1。
- 按一下「繼續」。

  在「交叉變數規則」索引標籤中，會自動選擇新定義的規則。
- 按一下「確定」。

交叉變數規則

圖表 7-25
交叉變數規則

規則	違反數目	規則表示式
Died Twice	27	(doa=1) & (dhosp=1)

交叉變數規律摘要列出至少違規一次的交叉變數規則、發生違規數和每個違規規則的說明。

觀察值報告

圖表 7-26
觀察值報告

	驗證規則違反		識別符		
觀察值	單一變數[a]	交叉變數	hospid	patid	physid
20		Died Twice	PBW	1192970826	355184
49		Died Twice	NHV	8717862852	237418
129		Died Twice	QWS	6901932085	215041
138		Died Twice	RLD	1205005069	695521
162		Died Twice	OZN	5546809538	125304
175	0 to 1 Dichotomy (1)		OZN	0333204686	883285
274	0 to 1 Dichotomy (1)		OZN	1038840465	103254
310	Nonnegative integer (1)		OZN	2090290204	883285
414		Died Twice	WPA	3351107142	462020
437	0 to 1 Dichotomy (1)		WPA	2349729006	723384
447		Died Twice	WPA	7163481282	519548
458		Died Twice	WPA	9159094175	652070
462		Died Twice	WPA	2137520354	723384
537		Died Twice	SLB	5246122506	928076
544		Died Twice	SLB	1605957462	506108
620		Died Twice	GFG	8141858966	828754
629		Died Twice	GFG	3397891610	539412
630		Died Twice	GFG	3397891610	539412
639		Died Twice	GFG	3962622031	327422
644		Died Twice	GFG	4271782383	749432
649		Died Twice	GFG	0950686750	618069
653		Died Twice	GFG	0663642766	001448
722		Died Twice	GFG	0418125590	877354
748		Died Twice	GFG	8744721380	539412
752	Nonnegative integer (1)		GFG	4993307441	828754

a. 違反規則的變數數量會出現在每個規則之後。

觀察值報告現在包括交叉變數規則違規的觀察值，以及先前發現單一變數規則違規的觀察值。這些觀察值全部需要報告到資料輸入中，以進行改正。

摘要

分析人員現在擁有要對資料輸入主管作初步報告所必需的資訊。

相關程序

「驗證資料」程序是有用的資料品質控制工具。

- 識別異常觀察值 程序會分析您的資料樣式，並識別具有一些隨類型而異的顯著值之觀察值。

章節
8

自動資料準備

準備資料以供分析是任何專案中最重要的步驟之一—也是傳統上最耗時的步驟之一。「自動資料準備」(ADP) 可為您處理工作、分析您的資料並識別修正、篩選出有問題或可能無用的欄位、在適當時衍生新屬性，以及透過智慧型篩選技術增進效能。您可以全**自動**方式使用演算法，以允許其選擇並套用修正，或以**互動**方式使用演算法，以在進行變更前先行預覽，然後視需要接受或拒絕變更。

使用 ADP 可讓您快速、輕鬆地準備資料以建立模式，不需事先了解統計相關概念。模式將可更快地建立並進行資料評分，此外，使用 ADP 可提高自動建立模式程序。

以互動方式使用自動資料準備

某資源有限的保險公司，打算調查屋主的保險理賠，希望建立標示可疑潛在詐欺理賠的模式他們有一個關於之前理賠的資訊樣本，收集於 insurance_claims.sav 中。 建立模式之前，他們會使用自動資料準備來準備建模用的資料。由於他們希望在套用轉換前檢閱提議的轉換，因此會在互動式模式使用自動資料準備。

選擇目標

▶ 若要以互動方式執行「自動資料準備」，從功能表中選擇：
轉換(T) > 準備建模用的資料 > 互動式(N)...

圖表 8-1
「目標」索引標籤

第一個索引標籤要求控制預設設定的目標，但目標之間的實際差異為何？藉由使用每個目標來執行程序，我們可以看到結果有哪些不同之處。

- 請確定已經選取「**權衡速度與準確度**」，然後按一下「**分析**」。

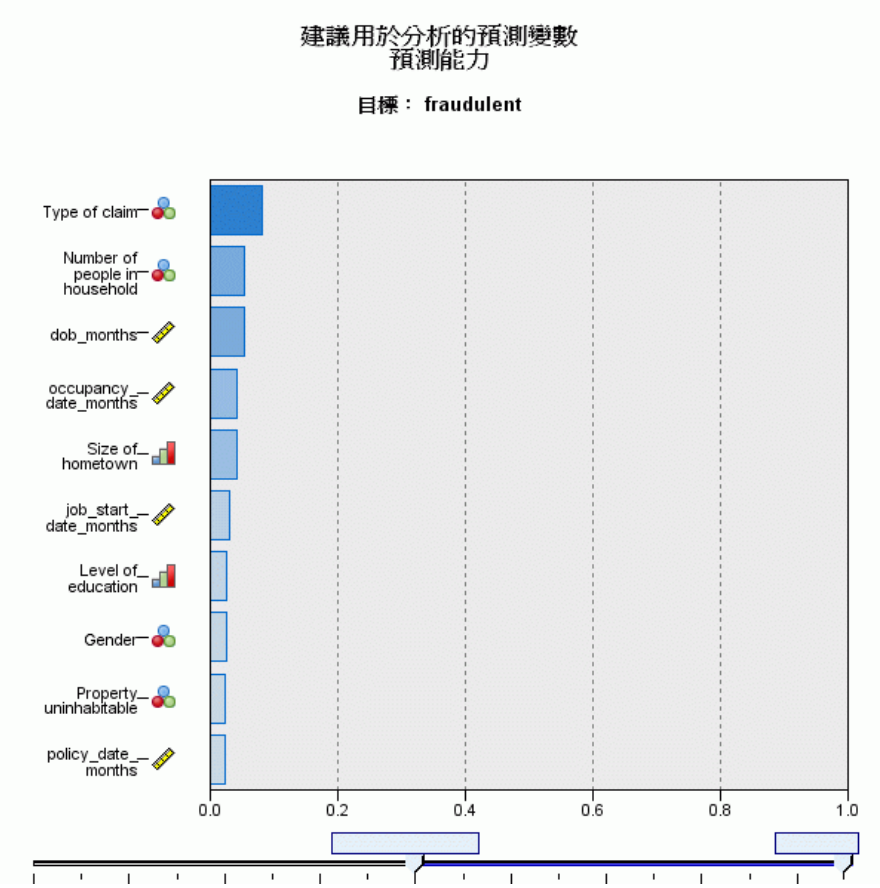
圖表 8-2
分析索引標籤，權衡目標的欄位處理摘要

欄位處理摘要

欄位		N
目標		1
預測變數		18
建議用於分析的預測變數	總數	18
	原始欄位 (未轉換)	9
	原始欄位轉換	4
	衍生自日期和時間	5
	已建構	0
未使用預測變數		0

當程序在處理資料時，焦點會自動切換至「分析」索引標籤。預設的主檢視是「欄位處理摘要」，能讓您一覽自動資料準備處理欄位的方式。建議用於建模的有一個目標、18 個輸入和 18 個欄位。在建議用於建模的欄位之中，有 9 個是原始的輸入欄位，4 個是原始輸入欄位的轉換，還有 5 個是衍生自日期和時間欄位。

圖表 8-3
分析索引標籤，權衡目標的預測能力



預設的輔助檢視屬於「預測能力」，這能快速讓您瞭解建模時哪些建議欄位最有用。請注意，雖然建議用於分析的預測值有 18 個，但在預設情況下，預測能力圖只會顯示前 10 個。若要顯示更多或更少欄位，請使用圖表下方的滑軸控制。

以「**權衡速度與準確度**」作為目標時，「理賠類型」即視為「最佳」預測值，接著是「家中人口數」和索賠者當時以月份為單位的年齡（計算從出生到當時日期的時間）。

- 按一下「**清除分析**」，再按一下「目標」索引標籤。
- 選取「**最佳化速度**」，並按一下「**分析**」。

圖表 8-4
分析索引標籤，最佳化速度時的欄位處理摘要

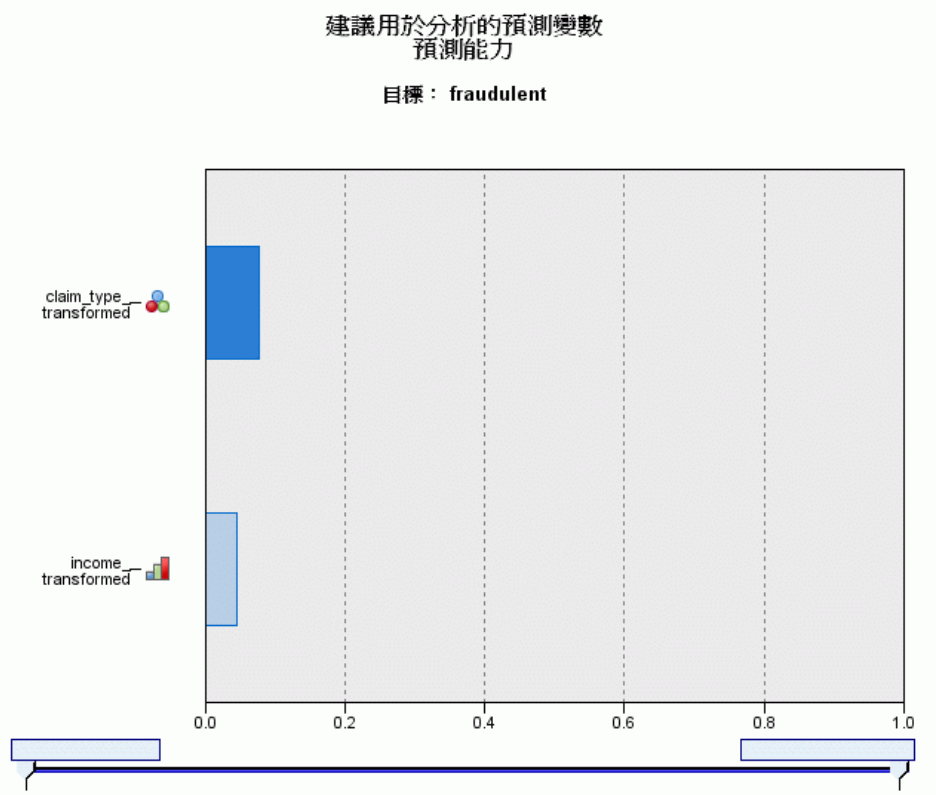
欄位處理摘要

欄位		N
目標		1
預測變數		18
建議用於分析的預測變數	總數	2
	原始欄位 (未轉換)	0
	原始欄位轉換	2
	衍生自日期和時間	0
	已建構	0
未使用預測變數		16

- 無法建構可用的預測變數。最常見的原因包括：與目標高度關聯的連續預測變數太少，或者所有連續預測變數都互不關聯。

當程序在處理資料時，焦點會再次自動切換至「分析」索引標籤。在此狀況下，建模用的建議欄位只有 2 個，而這兩個欄位都是原始欄位的轉換。

圖表 8-5
分析索引標籤，最佳化速度時的預測能力



以「**最佳化速度**」作為目標時，claim_type_transformed 即視為「最佳」預測值，接著是 income_transformed。

- 按一下「**清除分析**」，再按一下「目標」索引標籤。
- 選取「**最佳化準確度**」，並按一下「**分析**」。

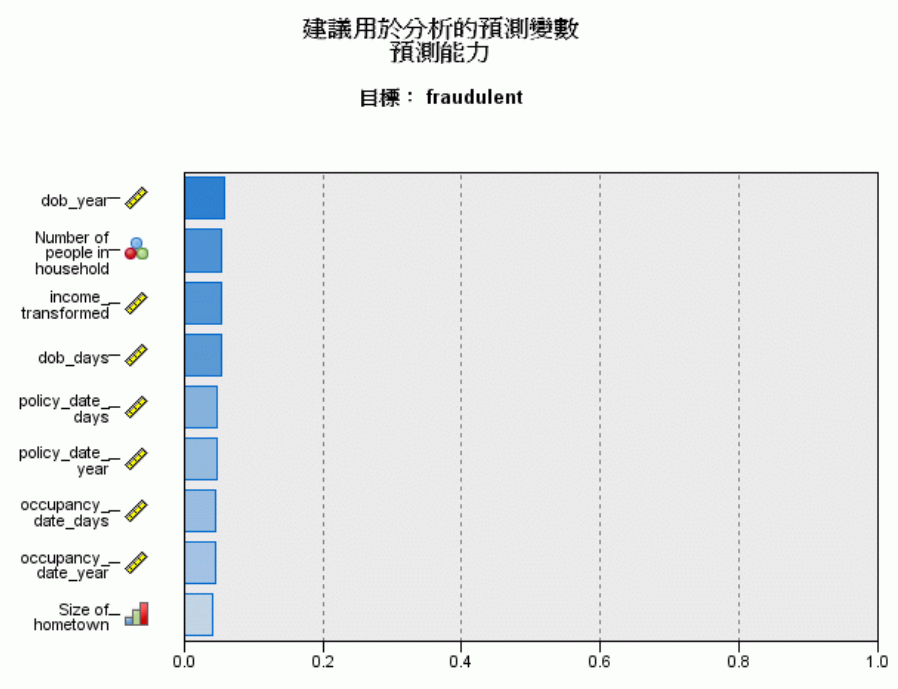
圖表 8-6
分析索引標籤，最佳化準確度時的預測能力

欄位處理摘要

欄位		N
目標		1
預測變數		18
建議用於分析的預測變數	總數	32
	原始欄位(未轉換)	9
	原始欄位轉換	4
	衍生自日期和時間	19
	已建構	0
未使用預測變數		0

以「**最佳化準確度**」作為目標時，建模用的建議欄位有 32 個，因為萃取日期、月份所產生的日期和時間，以及日期所產生的年份，還有時間所產生的時、分和秒，這些都衍生出更多的欄位。

圖表 8-7
分析索引標籤，最佳化準確度時的預測能力



將「理賠類型」視為「最佳」預測值，接著是索賠者開始從事最近一份工作的天數（計算工作起始日期到目前日期的時間），以及索賠者從事目前工作的年份（萃取自工作起始日期）。

若要摘要：

- 「**權衡速度與準確度**」會從日期中建立可用於建模的欄位，並且可能會轉換連續欄位（如「reside」）使其更偏向常態分配。
- 「**最佳化準確度**」會從日期建立部分額外的欄位（系統也會檢查偏離值，並且如果目標是連續的，可能會進行轉換使其更偏向常態分配）。
- 「**最佳化速度**」並不會準備日期，也不會重新調整連續欄位，但是當目標為類別時，會將類別預測值與 bin 連續預測值的類別合併（當目標為連續時，會執行功能選擇和建構）。

這家保險公司決定進一步探索「**最佳化準確度**」的結果。

▶ 從主檢視下拉式清單中選取「「**欄位**」。

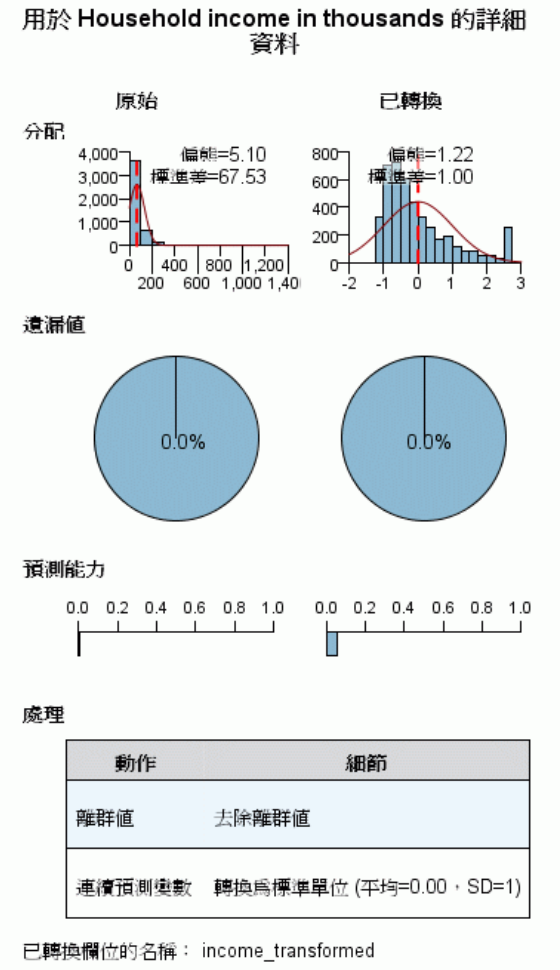
欄位和欄位詳細資料

圖表 8-8
欄位

「欄位」檢視顯示處理的欄位，以及 ADP 是否建議將它們用於建模。在任何欄位名稱上按一下可以在連結的檢視中顯示欄位的相關資訊。

- 按一下「收入」。

圖表 8-9
家庭收入（千元）的欄位詳細資料

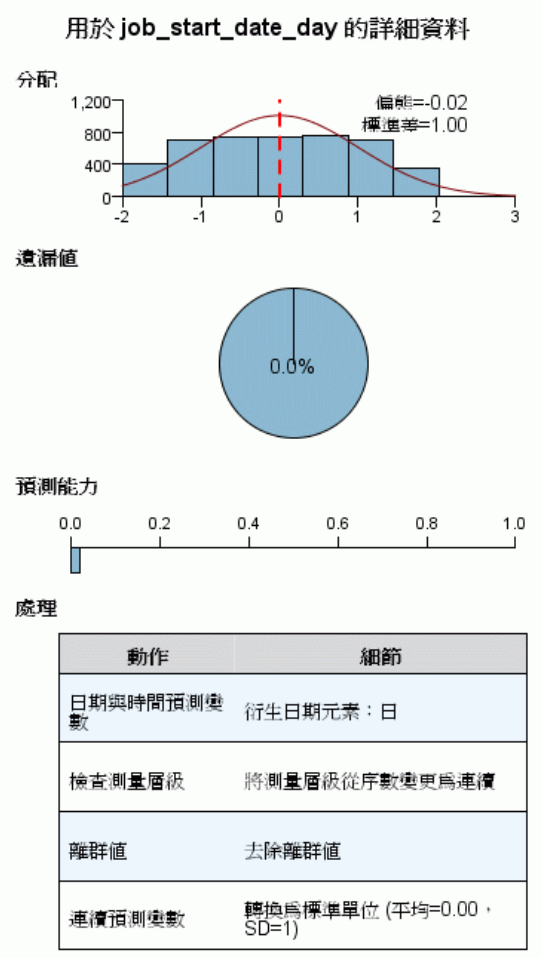


「欄位詳細資料」檢視會顯示原始和轉換後「家庭收入（千元）」的分配。根據處理表格，系統會將刪除視為偏離值的記錄（將其值設定為分割值以決定偏離值），並將欄位標準化，使其平均數為 0 而標準差為 1。而轉換欄位所產生的直方圖的最右邊有一個「突出物」，顯示系統視某一些記錄為偏離值，數目可能超過 200。收入的分配嚴重偏斜，所以這有可能是預設分割值在決定偏離值時太過積極的情況。

亦請注意，轉換欄位預測能力的增加超過原始欄位。這似乎是一個相當實用的轉換。

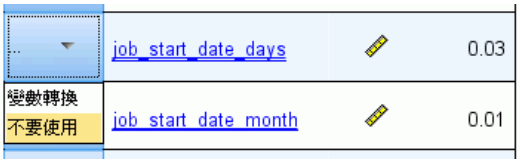
▶ 在「欄位」檢視中，按一下 job_start_date_day。（請注意，這不同於 job_start_date_days。）

圖表 8-10
job_start_date_day 的欄位詳細資料



欄位 job_start_date_day 是萃取自「雇用起始日 [job_start_date]」的日期。對於理賠是否為詐欺，這個欄位極不可能會有任何實際的影響，因此保險公司想將其移除，建模時不予考量。

圖表 8-11
家庭收入（千元）的欄位詳細資料



- 在「欄位」檢視中，於 job_start_date_day 列的「使用版本」下拉式清單中選取「**不使用**」。對所有字尾為 _day and _month 的欄位執行相同的作業。
- 若要套用轉換，請按一下「**執行**」。

現在建模用的資料集已經準備就緒，也就是說，所有建議預測值（不論新舊）的角色都已設定為「輸入」，而非建議預測值的角色則設定為「無」。若要建立僅有建議預測值的資料集，請使用對話方塊中的「套用轉換」設定。

以自動方式使用自動資料準備

某汽車業集團會追蹤各種個人汽車的銷售額。為了能夠識別表現超前與表現不佳的模式，您希望建立汽車銷售額與汽車特性之間的關係。此資訊收集於 car_sales_unprepared.sav 中。 使用自動資料準備以準備分析用的資料。同時使用準備「之前」和「之後」的資料來建模，以方便您比較結果。

準備資料

- 若要以自動模式執行自動資料準備，從功能表中選擇：
  轉換(T) > 準備建模用的資料 > 自動式(A)...

圖表 8-12
「目標」索引標籤

▶ 選取「**最佳化準確度**」。

因為目標欄位「銷售額（千元）」是連續欄位，並且可以在自動資料準備期間轉換，您想將轉換存成 XML 檔，以便能夠使用「反向轉換分數」對話方塊將轉換目標的預測值轉換回原始的尺度。

▶ 按一下「**設定**」索引標籤，再按一下「**套用並儲存**」設定。

圖表 8-13
「套用並儲存」設定

- 選取「將轉換儲存為 XML」，並按一下「瀏覽」瀏覽至 workingDirectory/car_sales_transformations.xml，用您想要儲存檔案的路徑取代 workingDirectory。
- 按一下「執行」。

這些選擇會產生下列指令語法：

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
```

```
 /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
 /REPLACEMISSING INPUT=YES TARGET=NO
 /REORDERNOMINAL INPUT=YES TARGET=NO
 /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
 /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
   CONSTRUCTION=NO
 /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
 /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
 /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
   MODE=FORWARD (ROLES=UPDATE)
 /SAVE TRANSFORMED=YES.
EXECUTE.
```

- ADP 指令會準備目標欄位 sales，然後透過 mpg 準備輸入欄位 resale 。
- PREPDATETIME 次指令已指定，但因為沒有欄位是日期或時間欄位，所以沒有使用。
- ADJUSTLEVEL 次指令會重新分配擁有 10 個值以上的次序欄位為連續欄位，以及擁有 5 個值以下的連續欄位為次序欄位。
- OUTLIERHANDLING 次指令會將距離平均數 3 個標準差以上的連續輸入（非目標）值，取代成距離平均數 3 個標準差的值。
- REPLACEMISSING 次指令會取代遺漏的輸入（非目標）值。
- REORDERNOMINAL 次指令會重新編碼名義輸入的值，從最不常出現到最常出現。
- RESCALE 次指令使用 z 分數轉換將連續輸入標準化，使其平均數為 0 而標準差為 1，然後使用 Box-Cox 轉換將連續目標標準化，使其平均數為 0 而標準差為 1。
- TRANSFORM 次指令會關閉此次指令所指定的所有預設作業。
- CRITERIA 次指令可指定目標和輸入轉換的預設字尾。
- OUTFILE 次指令指定轉換應儲存至 /workingDirectory/car_sales_transformations.xml，此處的 /workingDirectory 是您欲儲存 car_sales_transformations.xml 的路徑。
- TMS IMPORT 指令會讀取 car_sales_transformations.xml 中的轉換，並套用至作用中的資料集，同時更新所轉換現有欄位的角色。
- EXECUTE 指令會使系統處理轉換。當此指令的使用為較長語法流的一部份時，您可以移除 EXECUTE 指令以節省部分處理時間。

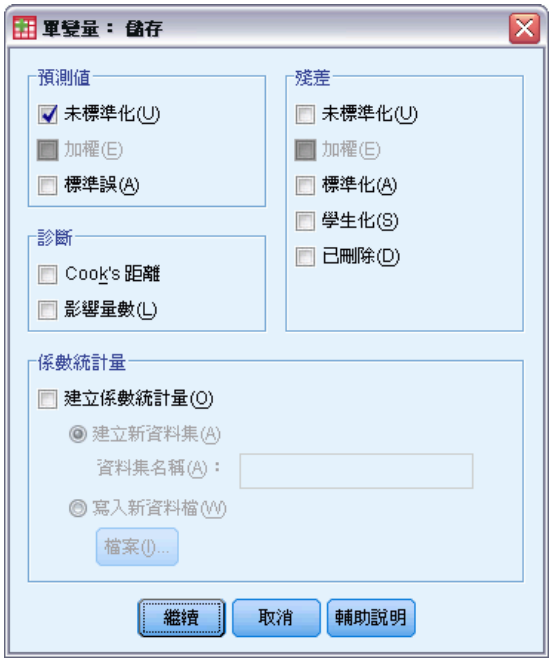
未準備資料的建模

▶ 若要建立未準備資料的模式，從功能表中選擇：
分析(A) > 一般線性模式 > 單變量...

圖表 8-14
「GLM 單變量」對話方塊

- 選取「銷售額（千元）[sales]」作為依變數。
- 選取「車輛類型 [type]」作為固定因子。
- 選取「4 年重新銷售值 [resale]」到「燃料效率 [mpg]」作為共變量。
- 按一下「儲存」。

圖表 8-15
「儲存」對話方塊



- 選取「預測值」組別中的「未標準化」。
- 按一下「繼續」。
- 按一下「GLM 單變量」對話方塊中的「確定」。

這些選擇會產生下列指令語法：

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

圖表 8-16
以未準備資料為基礎之模型的受試者間效應

依變數:Sales in thousands

來源	型 III 平方和	df	平均平方和	F	顯著性
校正後的模式	226123.658[a]	11	20556.696	5.050	.000
截距	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
誤差	427402.183	105	4070.497		
總數	1062354.955	117			
校正後的總數	653525.841	116			

a. R 平方 = .346 (調過後的 R 平方 = .277)

預設的「GLM 單變量」輸出包括受試者間效應，此為變異數表格分析。模式中的每個項目以及模式本身將整個接受測試，以檢驗其說明依變數變異的能力。請注意，此表格中並不會顯示變數標記。

預測值顯示不同的顯著水準；顯著值低於 0.05 的預測值一般都認為是對模式有用的。

準備資料的建模

圖表 8-17
「GLM 單變量」對話方塊

- 若要建立準備資料的模式，請叫回「GLM 單變量」對話方塊。
- 取消選取「銷售額（千元）[sales]」，並選取 sales_transformed 作為依變數。
- 取消選取「4 年重新銷售值 [resale]」到「燃料效率 [mpg]」，並選取 resale_transformed 到 mpg_transformed 作為共變量。
- 按一下「確定」。

這些選擇會產生下列指令語法：

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
   engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
   length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
   wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
   fuel_cap_transformed mpg_transformed type.
```

圖表 8-18
以準備資料為基礎之模型的受試者間效應

依變數:sales_transformed

來源	型 III 平方和	df	平均平方和	F	顯著性
校正後的模式	79.327[a]	11	7.212	13.638	.000
截距	2.436	1	2.436	4.606	.034
resale_transformed	.954	1	.954	1.804	.181
price_transformed	9.271	1	9.271	17.533	.000
engine_s_transformed	2.885	1	2.885	5.456	.021
horsepow_transformed	.034	1	.034	.064	.801
wheelbas_transformed	1.213	1	1.213	2.293	.132
width_transformed	.037	1	.037	.071	.791
length_transformed	.265	1	.265	.501	.480
curb_wgt_transformed	.103	1	.103	.194	.660
fuel_cap_transformed	.132	1	.132	.249	.618
mpg_transformed	3.390	1	3.390	6.411	.012
type	4.007	1	4.007	7.579	.007
誤差	76.673	145	.529		
總數	156.000	157			
校正後的總數	156.000	156			

a. R 平方 = .509 (調過後的 R 平方 = .471)

以未準備資料和準備資料所建立的模式中，其受試者間效應有幾點相當有趣的差異需要注意。首先，您會注意到總自由度增加；這是因為在自動資料準備期間，插補值取代遺漏值的緣故，這樣從第一個模式中完全移除的記錄可用於第二個模式。或許更值得注意的是，某些預測值的顯著性已經改變。當兩個模式都同意引擎大小 [engine_s] 和車輛類型 [type] 都對模式有用，而軸距 [wheelbas] 和空車重量 [curb_wgt] 已不再顯著時，現在車輛價格 [price_transformed] 和燃料效率 [mpg_transformed] 便會顯著。

為什麼會發生這種改變？「銷售額」有些偏斜的分配，所以有可能當銷售額轉換之後，原本有些影響力記錄的軸距和空車重量，便不再具有影響力。另一個可能性是，因為遺漏值取代而能夠使用的額外觀察值，改變了這些變數的統計顯著性。不論何種情況，都需要進一步調查，因此這個階段我們並不會購買。

請注意，根據準備資料所建立之模式的「R 平方」較高，但因為銷售額已經轉換，所以這可能不是用於比較每個模式效能的最佳量數。因此，您可以改為計算觀察值與兩組預測值之間的無母數相關性。

比較預測值

▶ 若要從兩個模式中取得預測值的相關性，從功能表中選擇：
分析(A) > 相關 > 雙變數...

圖表 8-19
「雙變數相關分析」對話方塊

- 選取「銷售額（千元）[sales]」、「銷售額預測值 [PRE_1]」和「轉換的銷售額預測值 [PRE_2]」作為分析變數。
- 取消選取「個人」，並選取「相關係數」組別中的「Kendall' s tau-b」和「Spearman」。

  請注意，「轉換的銷售額預測值 [PRE_2]」可用來計算無母數相關性，不需要反向轉換回原始尺度，因為反向轉換並不會變更預測值的等級順序。
- 按一下「確定」。

  這些選擇會產生下列指令語法：

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

圖表 8-20
無母數相關性

			Sales in thousands	Sales 的預測值	Sales_transformed 的預測值
Kendall's tau_b 統計量數	Sales in thousands	相關係數	1.000	.376**	.484**
		顯著性 (雙尾)	.	.000	.000
		個數	157	117	157
	Sales 的預測值	相關係數	.376**	1.000	.655**
		顯著性 (雙尾)	.000	.	.000
		個數	117	117	117
	Sales_transformed 的預測值	相關係數	.484**	.655**	1.000
		顯著性 (雙尾)	.000	.000	.
		個數	157	117	157
Spearman's rho 係數	Sales in thousands	相關係數	1.000	.530**	.666**
		顯著性 (雙尾)	.	.000	.000
		個數	157	117	157
	Sales 的預測值	相關係數	.530**	1.000	.831**
		顯著性 (雙尾)	.000	.	.000
		個數	117	117	117
	Sales_transformed 的預測值	相關係數	.666**	.831**	1.000
		顯著性 (雙尾)	.000	.000	.
		個數	157	117	157

**. 相關的顯著水準爲 0.01 (雙尾)。

第一行顯示使用準備資料所建立之模式的預測值，在 Kendall’s tau-b 和 Spearman’s rho 的量數方面，都與觀察值強烈相關。這個情況顯示執行自動資料準備已經改善了模式。

反向轉換預測值

- 準備資料包括銷售額的轉換，因此這個模式中的預測值並不剛好像分數那麼實用。若要將預測值轉換成原始尺度，從功能表中選擇：

  轉換(T) > 準備建模用的資料 > 反向轉換分數(B)...

圖表 8-21
「反向轉換分數」對話方塊

- 選取「轉換的銷售額預測值 [PRE_2]」作為要反向轉換的欄位。
- 輸入 _backtransformed 作為新欄位的字尾。
- 輸入 workingDirectory\car_sales_transformations.xml 取代 workingDirectory 中的檔案路徑，作為含有轉換的 XML 檔的位置。
- 按一下「確定」。

這些選擇會產生下列指令語法：

```
TMS IMPORT
 /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
   MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- TMS IMPORT 指令會讀取 car_sales_transformations.xml 中的轉換，並將反向轉換套用至 PRE_2。
- 含有反向轉換值的新欄位命名為 PRE_2_backtransformed。
- EXECUTE 指令會使系統處理轉換。當此指令的使用為較長語法流的一部份時，您可以移除 EXECUTE 指令以節省部分處理時間。

摘要

使用自動資料準備時，您可以快速取得可改善模式的資料轉換。如果目標轉換，您可以將轉換存成 XML 檔，並使用「反向轉換分數」對話方塊，將轉換目標的預測值轉換回原始尺度。

章節
9

識別特殊觀察值

「異常偵測」程序會搜尋以其集群標準的差異為基礎的異常觀察值。這個程序設計來以資料稽核為目的，在探索資料分析的步驟中，以及在任何推論資料分析前，快速偵測異常觀察值。這個演算法是為了一般異常偵測而設計；也就是異常觀察值的定義並非指定為任何特定的應用，例如在醫療保健產業中偵測異常付款模式或在金融產業中偵測洗錢，這些情況中可以完整定義一項異常狀況。

識別異常觀察值演算法

本演算法分為三個階段：

模式建立。 本程序建立資料集中用來說明自然分組（或集群）的集群模式，否則資料集中的自然分組將不明顯。集群以一組輸入變數為基礎。所得的集群模式，以及用來計算集群組標準的足夠統計量將儲存起來供日後使用。

計分。該模式套用至每個案例以找出其集群組，且建立一些各案例的指標以測量出該案例對於其集群組是否有不尋常之處。所有案例均按異常指標值進行排序。案例清單的最主要部分視為異常集合。

推理。對於每個異常案例，變數按其對應的變數偏離指標而排序。最主要變數、其值以及對應的標準值表示案例視為異常的原因。

識別醫療資料庫的異常觀察值

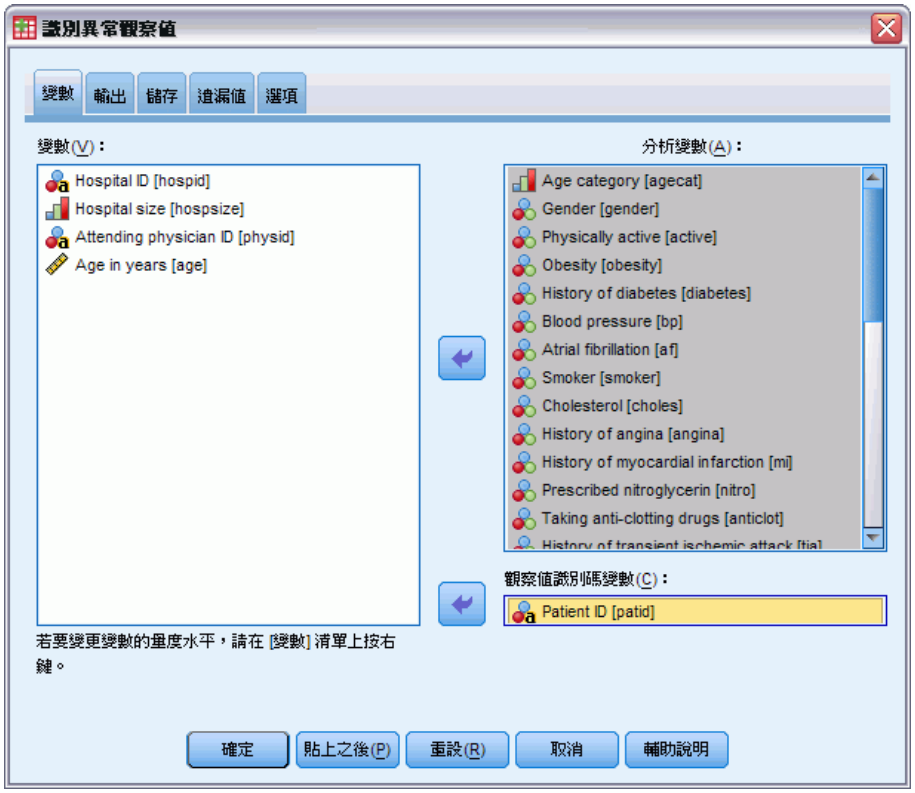
由於中風治療結果預測模型可能對異常觀察很敏感，因此受雇建立這些模型的資料分析人員很擔心資料品質。某些偏離的觀察值代表真正獨特的觀察值，因此不適合用來預測，然而其他由資料輸入錯誤所造成的觀察值，在技術上是「正確」的，因此而不會被資料驗證程序偵測到。

這個資訊收集於 stroke_valid.sav 中。 使用「識別異常觀察值」來清理資料檔案。可以在 detectanomaly_stroke.sps 找到重製這些分析的語法。

執行分析

▶ 若要識別異常觀察值，請從功能表選擇：
資料 > 識別特殊觀察值(I)...

圖表 9-1
「識別異常的觀察值」對話方塊，「變數」索引標籤



- 選取「年齡類別」到「中風 3 到 6 個月」做為分析變數。
- 選取「病人 ID」做為觀察值識別碼變數。
- 按一下「輸出」索引標籤。

圖表 9-2
「識別異常的觀察值」對話方塊，「輸出」索引標籤

- 選取「對等組別標準」、「異常指標」、「根據分析變數所發生的原因」，和「已處理的觀察值」。
- 按一下「儲存」索引標籤。

圖表 9-3
「識別異常的觀察值」對話方塊，「儲存」索引標籤

- 選取「異常指數」、「對等組別」、和「原因」。

  儲存這些原因，可以讓您產生將原因做成摘要之實用的散佈圖。

- 按一下「遺漏值」索引標籤。

圖表 9-4
「識別異常的觀察值」對話方塊，「遺漏值」索引標籤

- 選取「**在分析中包含遺漏值**」。這個程序是必要的，因為要處理在治療期間或治療之前死亡的病人，導致很多使用者遺漏值產生。分析中將新增一個作為尺度變數之額外變數，以測量每一觀察值中遺漏值的比例。
- 按一下「**選項**」索引標籤。

圖表 9-5
「識別異常的觀察值」對話方塊，「選項」索引標籤

- 輸入 2 以做為視為異常觀察值的百分比。
- 取消選取「只識別其異常指數值符合或超過最低值的觀察值」。
- 輸入 3 以做為原因的最大數量。
- 按一下「確定」。

觀察值處理摘要(O)

圖表 9-6
觀察值處理摘要(S)

		N	組合的 %	總數的 %
同級 ID	1	710	67.7%	67.7%
	2	90	8.6%	8.6%
	3	248	23.7%	23.7%
結合的		1048	100.0%	100.0%
總和		1048		100.0%

各個觀察值將歸類到相似的觀察值之對等組別中。觀察值程序摘要會顯示已建立對等組別的數量，以及在各個對等組別中觀察值的數量與百分比。

異常觀察值指數清單

圖表 9-7
異常觀察值指數清單

觀察值	patid	不規則索引
843	7840326167	2.837
510	0714726620	2.022
623	6553808330	2.014
501	6461046805	2.002
607	1077125669	1.897
884	2260043998	1.889
614	4030164769	1.869
241	1038840465	1.865
13	2191527525	1.826
172	4458028382	1.786
705	1336411777	1.778
651	4103977868	1.767
384	2247641363	1.767
839	0437454972	1.766
861	9746101913	1.757
19	7237535360	1.756
806	4391632997	1.756
871	6961938294	1.739
239	7315965190	1.738
887	6044244232	1.737
245	0816869249	1.736

異常指數是一種測量，反應關於其對等組別，不尋常的觀察值。會顯示 2% 之最高異常指數的觀察值，也會顯示觀察值的個數與 ID。會列出 21 個觀察值，數值範圍為 1.736 到 2.837。清單中第一個與第二個觀察值間的異常指數之數值相對上差異較大，顯示觀察值 843 可能為異常。其他觀察值必須逐項審查。

異常觀察值對等 ID 清單

圖表 9-8
異常觀察值對等 ID 清單

觀察值	patid	同級 ID	同級大小	同級大小百分比
843	7840326167	3	248	23.7%
510	0714726620	3	248	23.7%
623	6553808330	3	248	23.7%
501	6461046805	3	248	23.7%
607	1077125669	3	248	23.7%
884	2260043998	3	248	23.7%
614	4030164769	3	248	23.7%
241	1038840465	3	248	23.7%
13	2191527525	3	248	23.7%
172	4458028382	3	248	23.7%
705	1336411777	1	710	67.7%
651	4103977868	1	710	67.7%
384	2247641363	3	248	23.7%
839	0437454972	3	248	23.7%
861	9746101913	3	248	23.7%
19	7237535360	1	710	67.7%
806	4391632997	1	710	67.7%
871	6961938294	1	710	67.7%
239	7315965190	3	248	23.7%
887	6044244232	1	710	67.7%
245	0816869249	3	248	23.7%

可能的異常觀察值會搭配其對等組別成員資訊一起顯示。前 10 個觀察值（總計 15 個觀察值）屬於對等組別 3，其餘則屬於對等組別 1。

異常觀察值原因清單

圖表 9-9
異常觀察值原因清單

原因: 1

觀察值	patid	原因變數	變數衝擊	變數值	變數規範
843	7840326167	cost	.411	200.51	19.83
510	0714726620	cost	.120	96.59	19.83
623	6553808330	cost	.175	114.01	19.83
501	6461046805	barthel1	.084	80	(遺漏值)
607	1077125669	cost	.126	96.11	19.83
884	2260043998	cost	.138	99.73	19.83
614	4030164769	rankin1	.085	3	(遺漏值)
241	1038840465	barthel1	.115	25	(遺漏值)
13	2191527525	barthel1	.118	40	(遺漏值)
172	4458028382	barthel1	.120	100	(遺漏值)
705	1336411777	cost	.244	198.25	42.47
651	4103977868	barthel1	.064	30	95
384	2247641363	barthel1	.122	20	(遺漏值)
839	0437454972	barthel1	.109	95	(遺漏值)
861	9746101913	barthel1	.102	70	(遺漏值)
19	7237535360	barthel3	.080	5	100
806	4391632997	barthel2	.088	10	100
871	6961938294	barthel1	.094	5	95
239	7315965190	rankin1	.092	3	(遺漏值)
887	6044244232	stroke1	.066	1	0
245	0816869249	barthel1	.124	5	(遺漏值)

原因變數為最常促成使觀察值歸類為異常的變數。各異常觀察值的主要原因變數，會與其影響、該觀察值的數值與對等組別標準一起顯示。類別變數的對等組別標準（遺漏值）能表示對等組別中的多數觀察值含有變數的遺漏值。

變數影響統計量，就是原因變數對相同對等組別中觀察值離差影響的比例。將分析中的 38 個變數，包括遺漏比例變數一起計算，變數之期待的影響則為 1/38 = 0.026。觀察值 843 上之變數成本的影響是 0.411，此數值相當大。與對等組別 3 中觀察值的 19.83 之平均值互相比較，觀察值 843 之成本的數值是 200.51。

此對話方塊的選取項目會要求前三個原因的結果。

- 若要查看其他原因的結果，請連按兩下以啟動表格。
- 將「原因」從階層維度移至列維度。

圖表 9-10
異常觀察值原因清單（前 8 個觀察值）

觀察	原因	patid	原因變數	變數衝擊	變數值	變數規範
843	1	7840326167	cost	.411	200.51	19.83
	2	7840326167	barthel1	.076	65	(遺漏值)
	3	7840326167	rankin1	.044	2	(遺漏值)
510	1	0714726620	cost	.120	96.59	19.83
	2	0714726620	barthel1	.083	80	(遺漏值)
	3	0714726620	rehab	.068	3	(遺漏值)
623	1	6553808330	cost	.175	114.01	19.83
	2	6553808330	surgery	.089	2	(遺漏值)
	3	6553808330	barthel1	.089	70	(遺漏值)
501	1	6461046805	barthel1	.084	80	(遺漏值)
	2	6461046805	rehab	.068	3	(遺漏值)
	3	6461046805	rankin1	.063	1	(遺漏值)
607	1	1077125669	cost	.126	96.11	19.83
	2	1077125669	barthel1	.094	85	(遺漏值)
	3	1077125669	rehab	.072	3	(遺漏值)
884	1	2260043998	cost	.138	99.73	19.83
	2	2260043998	barthel1	.114	65	(遺漏值)
	3	2260043998	rehab	.072	3	(遺漏值)
614	1	4030164769	rankin1	.085	3	(遺漏值)
	2	4030164769	barthel1	.085	45	(遺漏值)
	3	4030164769	recbart1	.062	2	(遺漏值)

這個組態讓您更輕易地比較各個觀察值的前三個原因之相對貢獻。觀察值 843 有可能為異常，因為其成本的數值異常過大。相反的，沒有單一理由提供超過 0.10 給觀察值 501 的不尋常。

尺度變數標準

圖表 9-11
尺度變數標準

		同級 ID			結合的
		1	2	3	
Length of stay for rehabilitation	平均值	16.55	16.39	15.91	16.39
	標準 差	12.596	.000	6.834	10.887
Total treatment and rehabilitation costs in thousands	平均值	42.4673	3.5089	19.8273	33.7641
	標準 差	26.45401	.50997	20.17309	27.31266
遺漏比例	平均值	.006	.541	.354	.134
	標準 差	.021	.000	.083	.197

尺度變數標準會報告各個對等組別與整體之各個變數平均數與標準差。比較數值能提供一些指示，表示哪些變數有助於對等組別構成。

例如，所有三個對等組別中，康復期的時間長度之平均數都非常固定，表示此變數無助於對等組別構成。相反的，治療與康復的總成本（以千為單位）和遺漏比例都提供我們對於對等組別成員一些較深刻的瞭解。對等組別 1 含有最高平均成本與最少遺漏值。對等組別 2 含有非常低的成本與很多遺漏值。對等組別 3 含有中等的成本與遺漏值。

此組織顯示對等組別 2 由到達時已死亡的病人組成，因此產生極小的成本，並導致所有的治療與康復變數成為遺漏值。對等組別 3 可能包含很多在治療期間死亡的病人，因此僅產生治療成本卻不會產生康復成本，因而導致康復變數為遺漏值。對等組別 1 幾乎全由治療與康復期後存活的病人所組成，因此產生最高成本。

類別變數標準

圖表 9-12
類別變數標準（前 10 個變數）

		同級 ID			結合的
		1	2	3	
Age category	最受歡迎的類別	2	3	2	2
	頻率	277	25	81	383
	百分比	39.0%	27.8%	32.7%	36.5%
Gender	最受歡迎的類別	0	0	1	0
	頻率	361	46	126	529
	百分比	50.8%	51.1%	50.8%	50.5%
Physically active	最受歡迎的類別	1	0	0	0
	頻率	373	55	139	531
	百分比	52.5%	61.1%	56.0%	50.7%
Obesity	最受歡迎的類別	0	0	0	0
	頻率	555	67	178	800
	百分比	78.2%	74.4%	71.8%	76.3%
History of diabetes	最受歡迎的類別	0	0	0	0
	頻率	665	80	219	964
	百分比	93.7%	88.9%	88.3%	92.0%
Blood pressure	最受歡迎的類別	1	1	1	1
	頻率	445	49	139	633
	百分比	62.7%	54.4%	56.0%	60.4%
Atrial fibrillation	最受歡迎的類別	0	0	0	0
	頻率	641	83	216	940
	百分比	90.3%	92.2%	87.1%	89.7%
Smoker	最受歡迎的類別	0	0	0	0
	頻率	578	69	179	826
	百分比	81.4%	76.7%	72.2%	78.8%
Cholesterol	最受歡迎的類別	0	0	0	0
	頻率	406	52	136	594
	百分比	57.2%	57.8%	54.8%	56.7%
History of angina	最受歡迎的類別	0	0	0	0
	頻率	493	52	167	712
	百分比	69.4%	57.8%	67.3%	67.9%

類別變數標準與尺度標準目的大致是相同的，但是類別變數標準會報告典型的（最常用的）類別，以及落入此類別之對等組別中觀察值的個數與百分比。 比較數值會稍微比較難處理，例如，在第一眼看來，性別比抽煙者 較有助於集群構成，因為所有三個對等組別中抽煙者的典型類別是相同的，而對等組別 3 中性別的典型類別則為不同，因為性別只含有兩個數值，您可以推論對等組別 3 中有 49.2% 的觀察值含有數值 0，這與其他對等組別的百分比非常相似。相反的，抽煙者之百分比的範圍則為 72.2% 到 81.4%。

圖表 9-13
類別變數標準（已選取的變數）

類別變數	統計值	同級 ID 同級 ID 1	2	3	結合的
Dead on arrival	最受歡迎的類別	0	1	0	0
	頻率	710	90	248	958
	百分比	100.0%	100.0%	100.0%	91.4%
Initial Rankin score	最受歡迎的類別	0	(遺漏值)	5	5
	頻率	166	90	104	193
	百分比	23.4%	100.0%	41.9%	18.4%
CAT scan result	最受歡迎的類別	0	(遺漏值)	0	0
	頻率	607	90	184	791
	百分比	85.5%	100.0%	74.2%	75.5%
Clot-dissolving drugs	最受歡迎的類別	2	(遺漏值)	0	2
	頻率	318	90	129	394
	百分比	44.8%	100.0%	52.0%	37.6%
Died in hospital	最受歡迎的類別	0	(遺漏值)	1	0
	頻率	710	90	171	787
	百分比	100.0%	100.0%	69.0%	75.1%
Treatment result	最受歡迎的類別	1	(遺漏值)	1	1
	頻率	524	90	96	620
	百分比	73.8%	100.0%	38.7%	59.2%
Post-event preventative surgery	最受歡迎的類別	0	(遺漏值)	(遺漏值)	0
	頻率	323	90	171	369
	百分比	45.5%	100.0%	69.0%	35.2%
Post-event rehabilitation	最受歡迎的類別	0	(遺漏值)	(遺漏值)	0
	頻率	278	90	171	314
	百分比	39.2%	100.0%	69.0%	30.0%

由尺度標準所引起的疑慮可以在類別標準表格中加以確認。同輩組別 2 完全由到達時已死亡的病人組成，因此所有的治療與康復變數都為遺漏值。同輩組別 3 的大部分病人 (69.0%) 在治療期間死亡，因此康復變數的典型類別為 (遺漏值)。

異常指數摘要

圖表 9-14
異常指數摘要

	不規則清單中的數量	最小	最大	平均值	標準差
不規則索引	21	1.736	2.837	1.872	.240

不規則清單中的數量是由規格所決定: 不規則百分比是 2%

此表格提供在異常清單中觀察值之異常指數值的摘要統計量。

原因摘要

圖表 9-15
原因摘要（治療與康復變數）

	出現爲原因		變數衝擊統計量			
	頻率	百分比	最小	最大	平均值	標準差
Dead on arrival	0	.0%	.	.	.	.
Initial Rankin score	0	.0%	.	.	.	.
CAT scan result	0	.0%	.	.	.	.
Clot-dissolving drugs	0	.0%	.	.	.	.
Died in hospital	0	.0%	.	.	.	.
Treatment result	0	.0%	.	.	.	.
Post-event preventative surgery	0	.0%	.	.	.	.
Post-event rehabilitation	0	.0%	.	.	.	.
Rankin score at 1 month	0	.0%	.	.	.	.
Rankin score at 3 months	0	.0%	.	.	.	.
Rankin score at 6 months	0	.0%	.	.	.	.
Barthel index at 1 month	13	61.9%	.064	.124	.100	.021
Barthel index at 3 months	1	4.8%	.088	.088	.088	.
Barthel index at 6 months	1	4.8%	.080	.080	.080	.
Recoded Barthel index at 1 month	0	.0%	.	.	.	.
Recoded Barthel index at 3 months	0	.0%	.	.	.	.
Recoded Barthel index at 6 months	0	.0%	.	.	.	.
Stroke between release and 1 month	0	.0%	.	.	.	.
Stroke between 1 and 3 months	0	.0%	.	.	.	.
Stroke between 3 and 6 months	0	.0%	.	.	.	.
Length of stay for rehabilitation	0	.0%	.	.	.	.
Total treatment and rehabilitation costs in thousands	6	28.6%	.120	.411	.202	.112
遺漏比例	0	.0%	.	.	.	.
整體	21	100.0%	.064	.411	.127	.076

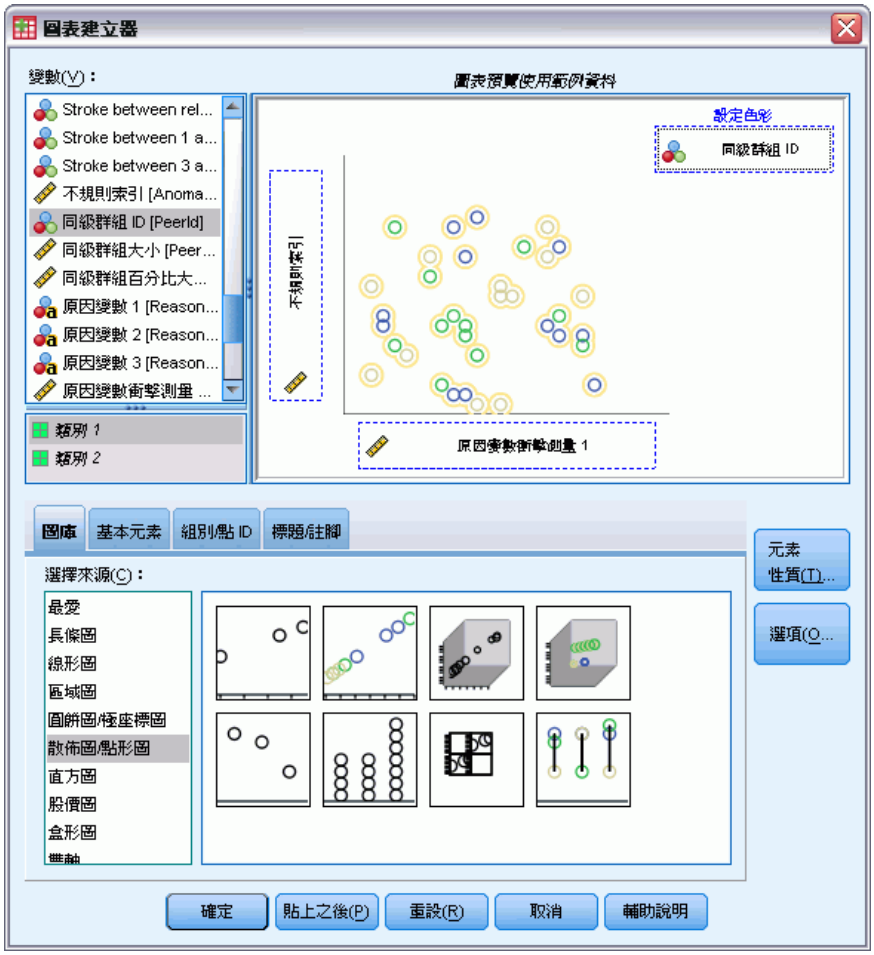
表格會針對分析中的每個變數，對其角色做成摘要，以做為主要原因。大部分的變數（例如從到達時已死亡到事故傷害後期康復中的變數）都不是任一觀察值存在異常清單中的主要原因。一個月的巴式量表 (Barthel Index) 為最常用的原因，接著是治療與康復的總成本（以千為單位）。變數影響統計量將進行摘要，並且會報告各個變數的最小值、最大值和平均數影響，以及各個變數（一個以上之觀察值的原因）的標準差。

根據變數影響之異常指數的散佈圖

此表格包含很多實用的資訊，但是可能很難掌握其關係。如果使用已儲存的變數，您可以建構讓這個過程更簡單的圖形。

▶ 若要產生散佈圖，請從功能表選擇：
統計圖(G) > 圖表建立器(C)...

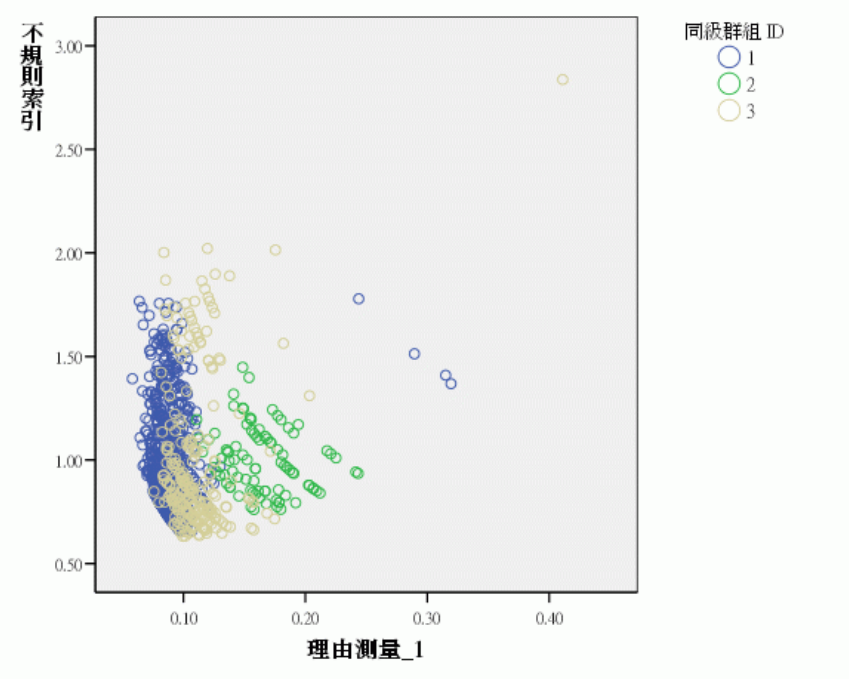
圖表 9-16
「圖表建立器」對話方塊



- 選取「**散佈圖/點形圖**」圖庫，並將「分組散佈圖」圖示拖曳至構圖區。
- 選取「異常指數」做為 y 變數，並選取「原因變數影響量數 1」做為 x 變數。
- 選取「對等組別 ID」作為用來設定顏色的變數。
- 按一下「**確定**」。

這些選擇會產生散佈圖。

圖表 9-17
第一原因變數作為影響量數的異常指數散佈圖



視察圖形能引導您進行一些觀察：

- 右上方的觀察值屬於對等組別 3，為最異常的觀察值，同時含有由單一變數提供最大貢獻的觀察值。
- 沿著 y 軸移動，我們看到有三個觀察值屬於組別 3，都含有剛好大於 2.00 的異常指數。這些觀察值應該視為異常而仔細加以檢驗。
- 沿著 x 軸移動，我們看到有四個觀察值屬於組別 1，其變數影響量數的範圍約為 0.23 到 0.33。這些觀察值應該更徹底地審查，因為這些數值將觀察值從散佈圖的點主體中分離出來。
- 對等組別 2 看來同質性很高，因為其異常指數和變數影響值與其集中趨勢並沒有很大的不同。

摘要

您已利用「識別異常觀察值」程序，標示出一些可供進一步檢查的觀察值。這些觀察值為無法由其他驗證程序識別者，因為這些變數（不只是變數本身的數值）之間的關係可以決定是否為異常觀察值。

僅大致根據兩個變數來建構對等組別，些許令人失望：到達時已死亡和在醫院中死亡。在進一步的分析中，您可以研究強迫建立較多對等組別的效應，或是您可以執行只包含治療後存活之病人的分析。

相關程序

「識別異常觀察值」程序是在資料檔案中偵測異常觀察值很實用的工具。

- 「驗證資料」程序可以識別可疑的與無效的觀察值、變數、以及作用中資料集的資料數值。

章節
10

最適 Binning

「Optimal Binning (最適 Binning)」程序可將各變數的數值分散成 bin，以離散化一個或多個尺度變數 (稱為「**Binning 輸入變數**」)。對「supervise (監督)」binning 處理的類別引導變數而言，Bin 資訊都是最適值。接著在需要或偏好使用類別變數的程序中進一步分析時，便可以使用 Bin，而非原始資料值。

最適 Binning 演算法

「最適 Binning」演算法的基本步驟特徵描述如下：

預先處理 (選用)。 Binning 輸入變數會分成 n 個 Bin (n 由您指定)，每個 Bin 包含相同的觀察值數目，或是儘可能包含相同的觀察值數目。

辨別可能的分割點。 在 Binning 輸入變數的不同值中，如果值與第二大的 Binning 輸入變數不同值之引導變數不屬於相同類別，則每個不同值皆為可能的分割點。

選取分割點。 MDLP 接受條件會評估可產生最大資訊增益的可能分割點。請重複步驟，直到不會接受任何可能分割點為止。已接受的分割點會定義 Bin 的端點。

使用最適 Binning 離散化貸款申請人資料

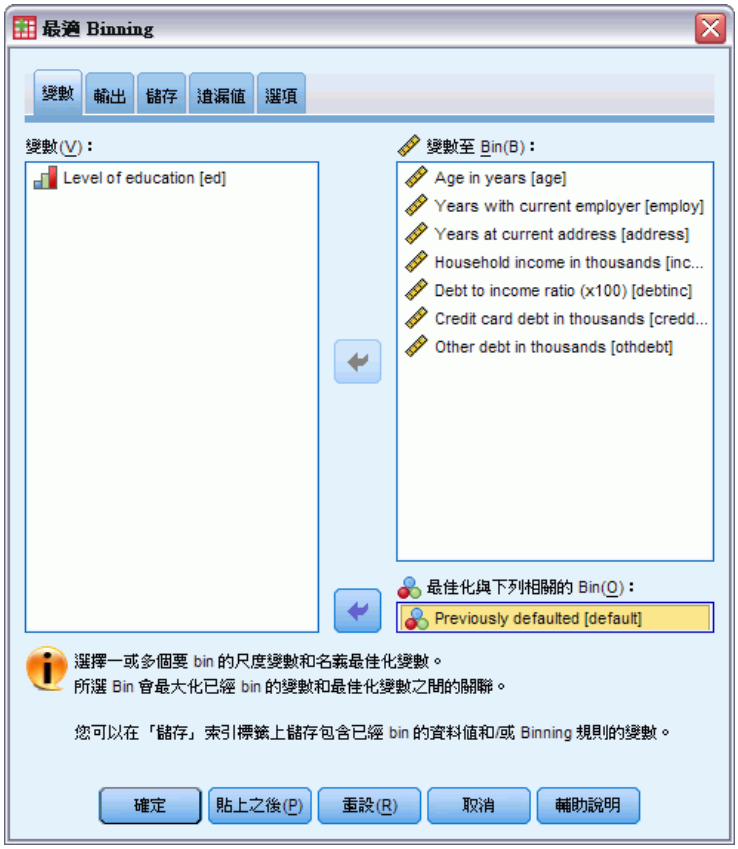
銀行的放貸人員努力降低貸款拖欠的比率，因此某放貸人員收集了過去與現在客戶的財務與人口資訊，希望能夠建立一個可預測貸款拖欠機率的模式。在可能的預測值中，有數個是尺度預測值，但放貸人員想要考量的是最適合類別預測值的模式。

5000 個過去客戶的資訊收集於 bankloan_binning.sav 中。 使用「最適 Binning」程序為尺度預測值產生 Binning 規則，然後使用產生的規則處理 bankloan.sav。已處理的資料集便可以用來建立預測模式。

執行分析

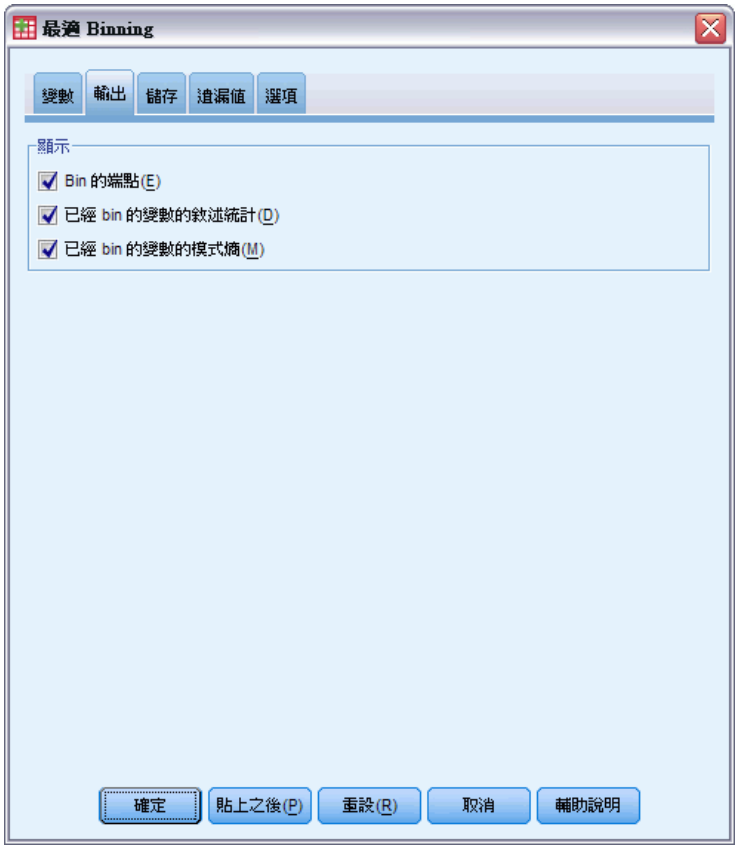
▶ 若要執行「最適 Binning」分析，請從功能表中選擇：
轉換 > 最適 Binning...

圖表 10-1
「最適 Binning」對話方塊，「變數」索引標籤



- 選取「年齡（年為單位）」與「服務於目前雇主的年數」到「其他負債（以千為單位）」作為進行 Bin 的變數。
- 選取「先前已拖欠」為引導變數。
- 按一下「輸出」索引標籤。

圖表 10-2
「最適 Binning」對話方塊，「輸出」索引標籤



- 為進行 Bin 的變數選取「敘述統計」與「模式熵」。
- 按一下「儲存」索引標籤。

圖表 10-3
「最適 Binning」對話方塊，「儲存」索引標籤

- 選取「建立包含已 bin 資料值的變數」。
- 輸入語法檔的路徑與檔案名稱，以包含產生的 Binning 規則。在本範例中，我們使用的是 /bankloan_binning-rules.sps。
- 按一下「確定」。

這些選擇會產生下列指令語法：

```
* Optimal Binning.
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
  othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bin
  creddebt_bin othdebt_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
  /OUTFILE RULES='/bankloan_binning-rules.sps'
  /PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- 本程序使用 MDLP binning，以「預設值」 binning 引導變數，將「年齡」、「雇用」、「地址」、「收入」、「debtinc」、「creddebt」、和「othdebt」 binning 輸入變數離散化。
- 這些變數的離散化值將儲存在新變數「age_bin」、「employ_bin」、「address_bin」、「income_bin」、「debtinc_bin」、「creddebt_bin」、和「othdebt_bin」中。
- 如果 binning 輸入變數有超過 1000 個不同數值，則在執行 MDLP binning 前會以相同次數方法將數目減少到 1000。
- 代表 binning 規則的指令語法儲存在 /bankloan_binning-rules.sps 檔中。
- Binning 輸入變數要求 bin 端點、敘述統計、和模式熵值。
- 其他 binning 條件皆設為其預設值。

敘述統計

圖表 10-4
敘述統計量

	個數	最小值	最大值	明確值個數	Bin 個數
Age in years	5000	20	58	39	2
Years with current employer	5000	0	38	39	4
Years at current address	5000	0	37	38	3
Household income in thousands	5000	12.10	2461.70	1100	2
Debt to income ratio (x100)	5000	.08	44.62	2060	5
Credit card debt in thousands	5000	.01	139.58	5000	4
Other debt in thousands	5000	.01	416.52	4999	2

敘述統計表提供 Binning 輸入變數的摘要資訊。前四欄是有關預先 Bin 的值。

- **N** 為分析中使用的觀察值數目。當使用刪除全部遺漏值處理方式時，任何一個變數的這個值應為常數。使用成對遺漏值處理方式時，這個值可能不是常數。由於這個資料集沒有遺漏值，這個值就是觀察值數目。
- 「**最小值**」與「**最大值**」欄顯示資料集中每個 Binning 輸入變數的 (預先 Binning) 最小值與最大值。除了瞭解所觀察到每個變數的值範圍外，這些值在找出預期範圍外的值時很有用。
- 「**不同值數目**」會告訴您已使用相同次數分配演算法預先處理了哪一些變數。依照預設值，具有超過 1000 個不同值 (從「家庭收入 (以千為單位)」到「其他負債 (以千為單位)」) 的變數已預先 Bin 處理到 1000 個不同的 Bin 中。接著系統會使用 MDLP，將這些預先處理的 Bin 對著引導變數進行 Bin 處理。您可以在「選項」索引標籤上控制預先處理功能。
- 「**Bin 數量**」是由程序產生的最後 Bin 數目，此數目遠小於不同值數目。

模式熵

圖表 10-5
模式熵

	模式熵
Age in years	.788
Years with current employer	.754
Years at current address	.781
Household income in thousands	.803
Debt to income ratio (x100)	.711
Credit card debt in thousands	.776
Other debt in thousands	.801

模式熵愈小，表示引導變數 Previously defaulted 的 Bin 變數預測準確性愈高。

模式熵能夠讓您知道在預測模式中，每個變數能夠預測拖欠機率的程度。

- 對於每個產生的 Bin 而言，最佳的可能預測值會包含與引導變數相同的值，因此可完全預測引導變數。這類預測值皆有未定義的模式熵。這是在真實世界中不會發生的狀況，如果發生了，表示資料的品質可能有問題。
- 最差的可能預測值是比猜測還糟的預測值，其模式熵的值會視資料而異。這個資料集中總共有 5000 個客戶，其中有 1256 (或 0.2512) 個客戶拖欠，3744 (或 0.7488) 個客戶沒有拖欠，因此最差的可能預測值之模式熵為 $-0.2512 \times \log_2(0.2512) - 0.7488 \times \log_2(0.7488) = 0.8132$。

除了具低模式熵值的變數應產生更佳預測值之外，我們很難做出更明確的陳述，因為一個好的模式熵值的構成因素是取決於應用程式與資料。在此狀況下，相對於不同類別的數目，所產生 Bin 數目較大的變數其模式熵值似乎較低。您應使用預測模式程序對這些作為預測值的 Binning 輸入變數執行進一步的評估，預測模式程序具有更多的變數選擇工具。

Binning 摘要

Binning 摘要會按照引導變數的值來報告已產生 Bin 的界限，與每個 Bin 的次數個數。系統會為每個 Binning 輸入變數產生不同的 Binning 摘要表格。

圖表 10-6
「年齡（年為單位）」的 Binning 摘要

	端點		水準 Previously defaulted 觀察值數量		
Bin	下界	上界	No	Yes	總數
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Age in years < 上界。

a. 無界限

「年齡（年為單位）」摘要顯示將 1768 個客戶（年齡皆為 32 歲或以下）放入 Bin 1，將其餘的 3232 個客戶（年齡皆超過 32 歲）放入 Bin 2。在 Bin 1 中先前拖欠的客戶比例 (639/1768=0.361) 遠高於 Bin 2 (617/3232=0.191)。

圖表 10-7
「家庭收入（以千為單位）」的 Binning 摘要

Bin	端點		水準 Previously defaulted 觀察值數量		
	下界	上界	No	Yes	總數
1	[a]	26.70	1054	513	1567
2	26.70	[a]	2690	743	3433
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Household income in thousands < 上界。

a. 無界限

「家庭收入（以千為單位）」摘要顯示類似的模式，有單一分割點 26.70，且 Bin 1 中先前拖欠的客戶比例 (513/1567=0.327) 比 Bin 2 (743/3433=0.216) 的高。如同從模式熵統計量所預期，這些比例中的差異不如「年齡（年為單位）」的差異大。

圖表 10-8
「其他負債（以千為單位）」的 Binning 摘要

Bin	端點		水準 Previously defaulted 觀察值數量		
	下界	上界	No	Yes	總數
1	[a]	2.19	2161	539	2700
2	2.19	[a]	1583	717	2300
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Other debt in thousands < 上界。

a. 無界限

「其他負債（以千為單位）」摘要顯示相反的模式，有單一分割點 2.19，而 Bin 1 中先前已拖欠的客戶比例 (539/2700=0.200) 比 Bin 2 (717/2300=0.312) 的低。同樣的，如同從模式熵統計量所預期，這些比例中的差異不如「年齡（年為單位）」的差異大。

圖表 10-9
「服務於目前雇主的年數」的 Binning 摘要

Bin	端點		水準 Previously defaulted 觀察值數量		
	下界	上界	No	Yes	總數
1	[a]	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Years with current employer < 上界。

a. 無界限

「服務於目前雇主的年數」摘要顯示拖欠者的比例隨著 Bin 數目的增加而下降。

Bin	拖欠者比例
1	0.432
2	0.302

Bin	拖欠者比例
3	0.154
4	0.078

圖表 10-10
「現址居住年數」的 Binning 摘要

	端點		水準 Previously defaulted 觀察值數量		
Bin	下界	上界	No	Yes	總數
1	[a]	7	1652	829	2481
2	7	14	1184	313	1497
3	14	[a]	908	114	1022
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Years at current address < 上界。

a. 無界限

「現址居住年數」摘要顯示類似的模式。如同從模式熵統計量所預期，「服務於目前雇主的年數」Bin 之間的差異（以拖欠者比例表示）比「現址居住年數」的大。

Bin	拖欠者比例
1	0.334
2	0.209
3	0.112

圖表 10-11
「信用卡負債（以千為單位）」的 Binning 摘要

	端點		水準 Previously defaulted 觀察值數量		
Bin	下界	上界	No	Yes	總數
1	[a]	.97	2169	466	2635
2	.97	1.91	848	307	1155
3	1.91	6.05	643	352	995
4	6.05	[a]	84	131	215
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Credit card debt in thousands < 上界。

a. 無界限

「信用卡負債（以千為單位）」摘要顯示相反的模式，拖欠者比例隨 Bin 數目的增加而增加。「服務於目前雇主的年數」和「現址居住年數」較能識別出準時還款機率高的人，「信用卡負債（以千為單位）」較能識別出拖欠機率高的人。

Bin	拖欠者比例
1	0.177
2	0.266
3	0.354
4	0.609

圖表 10-12
「負債與收入比率 (x100)」的 Binning 摘要

Bin	端點		水準 Previously defaulted 觀察值數量		
	下界	上界	No	Yes	總數
1	[a]	4.39	912	88	1000
2	4.39	12.09	2006	437	2443
3	12.09	18.71	625	386	1011
4	18.71	31.00	198	303	501
5	31.00	[a]	3	42	45
總數			3744	1256	5000

計算各個 Bin 的方式爲下界 <= Debt to income ratio (x100) < 上界。

a. 無界限

「負債與收入比率 (x100)」摘要顯示與「信用卡負債 (以千為單位)」類似的模式。這個變數的模式熵值最低，因此是拖欠機率的最佳準預測值。此變數比「信用卡負債 (以千為單位)」更能分類出拖欠機率高的人，且分類出拖欠機率低的人的能力幾乎與「服務於目前雇主的年數」一樣好。

Bin	拖欠者比例
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

Bin 變數

圖表 10-13
「資料編輯程式」中 bankloan_binning.sav 的 Bin 變數

	default	age_bin	employ_bin	address_bi	income_bin	debtinc_bin	creddebt_bi	othdebt_bin
1	0	2	3	2	2	2	1	2
2	0	1	3	2	2	3	2	2
3	0	2	3	3	2	2	3	2
4	0	2	3	3	2	4	3	2
5	0	2	2	3	1	3	2	2
6	0	2	1	2	2	1	1	1
7	1	2	1	1	1	3	2	1
8	0	2	4	2	2	3	2	2
9	0	2	3	2	2	2	2	2
10	0	2	2	2	2	2	2	2
11	0	1	1	1	1	2	1	1
12	1	2	3	2	2	4	4	2
13	0	2	3	3	2	2	3	2
14	1	2	3	1	2	2	1	1
15	0	1	1	2	2	2	2	1

資料檢視 變數檢視

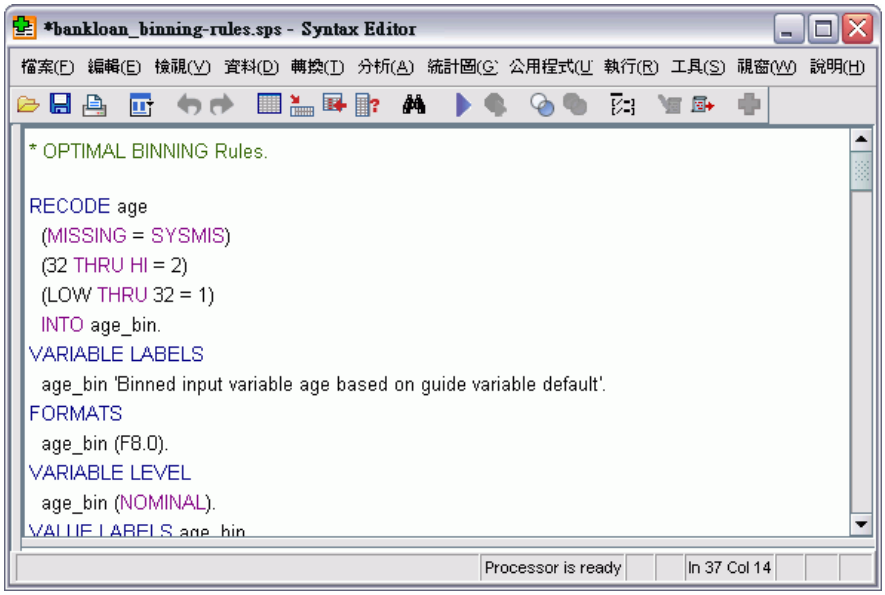
此資料集中 Binning 程序的結果在「資料編輯程式」中十分明顯。如果您要使用敘述程序或報告程序產生自訂的 Binning 結果摘要，這些 Bin 變數很有用，但不建議使用這個資料集來建立預測模式，因為 Binning 規則是使用這些觀察值所產生的。較佳的規劃是將 Binning 規則套用到另一個包含其他客戶資訊的資料集中。

套用語法 Binning 規則

執行「最適 Binning」程序時，您已要求將該程序產生的 Binning 規則儲存為指令語法。

- 開啟 bankloan_binning-rules.sps。

圖表 10-14
語法規則檔案



對於每個 Binning 輸入變數，會有一個指令語法區塊會執行 Binning；設定變數標記、格式與水準；設定 Bin 的數值標記。這些指令會套用到具有與 bankloan_binning.sav 相同變數的資料集。

- 開啟 bankloan.sav。
- 回到 bankloan_binning-rules.sps 的「語法編輯器」檢視。

▶ 若要套用 Binning 規則，請從「語法編輯器」功能表中選擇：
執行 > 全部...

圖表 10-15
「資料編輯程式」中 bankloan.sav 的 Bin 變數

	preddef3	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	.21304	2	3	2	2	2	4	2
2	.43690	1	3	1	2	3	2	2
3	.14102	2	3	3	2	2	1	1
4	.10442	2	3	3	2	1	3	1
5	.43690	1	1	1	2	3	2	2
6	.23358	2	2	1	1	2	1	1
7	.81709	2	4	2	2	4	3	2
8	.11336	2	3	2	2	1	1	1
9	.66390	1	2	1	1	4	2	2
10	.51553	2	1	2	1	4	3	1
11	.09055	1	1	1	1	1	1	1
12	.13631	1	2	1	1	2	1	1
13	.22890	2	4	3	2	2	3	2
14	.40484	2	2	2	2	3	2	2
15	.20866	2	4	3	2	2	3	2

資料檢視 變數檢視

系統已根據在 bankloan_binning.sav 上執行「最適 Binning」程序產生的規則，將 bankloan.sav 中的變數進行 Bin 處理。現在此資料集已備妥，可用於建立偏好使用或需要類別變數的預測模式。

摘要

我們已使用「最適 Binning」程序，針對為拖欠機率可能預測值的尺度變數產生了 Binning 規則，並將這些規則套用到個別的資料集。

在 Binning 程序期間，您會注意到經過 Bin 處理的「服務於目前雇主的年數」和「現址居住年數」較能識別出準時還款機率高的人，「信用卡負債 (以千為單位)」較能識別出拖欠機率高的人。當建立拖欠機率的預測模式時，這項有趣的觀察可提供給您其他的觀點。如果避免呆帳是主要的考量，則「信用卡負債 (以千為單位)」會比「服務於目前雇主的年數」與「現址居住年數」還要重要。如果以客戶數量的成長為優先考量，則「服務於目前雇主的年數」與「現址居住年數」較重要。

附錄
A

範例檔案

與產品同時安裝的範例檔存放在安裝目錄的範例子目錄中。在下列每種語言的「範例」子目錄中存有個別資料夾：英文、法文、德文、義大利文、日文、韓文、波蘭文、俄文、簡體中文、西班牙文和繁體中文。

並非所有範例檔案皆提供各種語言。如果範例檔案沒提供您需要的語言，語言資料夾有英文版的範例檔案。

說明

以下是使用於本文件中不同範例的範例檔之簡要描述。

- **accidents.sav。** 這是有關某保險公司研究年齡和性別風險因子對給定地區汽車意外事件的假設資料檔。每一個觀察值對應至一個年齡類別和性別的交叉分類。
- **adl.sav。** 這是有關致力於確定一個建議中風病患治療類型之效益的假設資料檔。醫師隨機指定女性中風病患至兩個組別之一。第一組接受標準的物理治療，而第二組則接受額外的情緒治療。在治療了三個月後，將每一個病患進行日常活動的能力記分為次序變數。
- **advert.sav。** 這是有關一家零售商致力於調查廣告費與廣告後銷售情形之間的關係的假設資料檔。為了這個目的，他們收集了過往銷售數字和相關的廣告費用。
- **aflatoxin.sav。** 這是有關檢定玉米作物是否有黃麴毒素（一種毒物，其濃度在介於和處於作物產量中都有很大的差異）的假設資料檔。一名穀物加工者收到來自 8 個作物產量各 16 個樣本，並以十億當量 (PPB) 來測量黃麴毒素的水準。
- **aflatoxin20.sav。** 這個資料檔包含由 aflatoxin.sav 取得，來自 4 和 8 作物產量的 16 個樣本，每一個樣本的黃麴毒素測量。
- **anorectic.sav。** 在將厭食/暴食行為症狀學標準化的過程中，研究人員 (Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997) 研究了 55 個飲食失調的青少年。每個病患在四年之中被訪問四個回合，所以得到總數為 220 的觀察值。在每次觀察中，為病患在 16 種症狀上逐一評分。目前遺漏了第二次訪察的病患 71，第二次訪察的病患 76，以及第三次訪察的病患 47 的症狀分數，因此只剩下 217 個有效觀察值。
- **autoaccidents.sav。** 這是有關一位保險分析師致力於為每個駕駛的汽車意外事件次數建立模式，同時考量駕駛的年齡和性別的假設資料檔。每一個觀察值代表一位不同的駕駛，記錄了駕駛的性別、年齡、和近五年內的汽車意外事故次數。
- **band.sav。** 本資料檔包含某樂團音樂 CD 假設性的每週銷售數字。也包含三個可能預測變數的資料。
- **bankloan.sav。** 這是有關一家銀行致力於減少放款利率預設值的假設資料檔。本檔包含 850 位以前的客戶與現在的準客戶的財務和人口資料。前 700 個觀察值為以前有借貸的客戶。最後 150 個觀察值是銀行需要作信用風險優良與不良分類的準客戶。

- **bankloan_binning.sav。** 這是包含 500 位以前客戶的財務和人口資料的假設資料檔。
- **behavior.sav。**在典型範例 (Price 和 Bouffard, 1974) 中，52 名學生被要求為 15 種情境與 15 種行為組合評等，等級共分為 10 點，從 0 =「非常適當」到 9 =「 非常不適當」。平均值超過個別值，值會被視為相異性。
- **behavior_ini.sav。** 本資料檔包含 behavior.sav 之二維解的起始組態。
- **brakes.sav。** 這是有關一間生產高性能汽車碟型煞車片工廠中品質管制的假設資料檔。資料檔包含由 8 個生產機器分別取得 16 個碟片的直徑測量。煞車的目標直徑是 322 公釐。
- **breakfast.sav。**在經典研究中(Green 和 Rao, 1972)，21 名 Wharton 學院 MBA 學生及其配偶被要求為 15 項早餐食品按喜愛程度分出等級：從 1 = 「最喜愛」到 15 =「最不喜愛」。他們的喜愛程度分六種不同情況記錄，從「整體喜愛」到「點心，僅配飲料」。
- **breakfast-overall.sav。** 本資料檔只包含第一種情況－「整體喜愛」－所喜愛的早餐項目。
- **broadband_1.sav。** 這是包含全國性寬頻服務地區用戶數目的假設資料檔。本資料檔包含四年期間 85 個地區每月的用戶數目。
- **broadband_2.sav。** 本資料檔與 broadband_1.sav 相同，但多了三個月的資料。
- **car_insurance_claims.sav。** 一個在別處 (McCullagh 和 Nelder, 1989) 出現和分析過，有關汽車損害理賠的資料集。理賠金額的平均數可建立模式為具有 gamma 分配，使用反連結函數將依變數的平均數相關至一被保險人年齡、車輛類型、和車齡的線性組合。提出理賠的數量可以用作尺度權重。
- **car_sales.sav。** 本資料檔包含假設性的銷售估計、定價、和不同的品牌與車輛型式的實體規格。定價和實體規格是由 edmunds.com 和製造商處輪流取得。
- **car_sales_uprepared.sav。** 這是 car_sales.sav 的修改版本，其中不包含任何欄位的轉換版本。
- **carpet.sav。**在一個普遍的範例 (Green 和 Wind, 1973) 中，計劃銷售全新地毯清潔機的公司想要檢驗影響消費者偏好的五個因子—包裝設計、品牌名稱、價格、「優秀家用品」獎章及退費保證。包裝設計有三個因子水準，每個水準中的清潔刷位置都不相同；三個品牌名稱 (K2R、Glory、及 Bissell)；三個價格水準；且最後兩個因子各有兩個水準 (無論無或有)。十名消費者將這些因子所定義的 22 種組合分級。「偏好」變數包含每個組合平均排名的等級。排名數值較小者會對應高偏好程度。這個變數反映每個組合偏好的整體量數。
- **carpet_prefs.sav。** 本資料檔是根據 carpet.sav 所描述的相同範例，但它包含 10 個消費者每一個人的實際等級。消費者被要求將 22 個產品組合從最喜歡排列到最不喜歡。變數「PREF1」到「PREF22」包含相關組合的識別碼，如 carpet_plan.sav 中所定義。
- **catalog.sav。** 本資料檔包含郵購公司銷售三項產品的每月假設銷售數字。也包含五個可能預測變數的資料。
- **catalog_seasfac.sav。** 本資料檔與 catalog.sav 相同，不過多了一組由「週期性分解」程序所計算的週期性因子以及隨附的資料變數。
- **cellular.sav。** 這是有關一家手機公司致力於減少顧客不忠的假設資料檔。顧客不忠傾向分數套用於帳戶，範圍由 0 至 100。帳戶分數 50 或以上有可能正尋求變更供應商。

- **ceramics.sav。** 這是有關一家製造商致力於確定一種新的優良合金是否較標準的合金有較大的耐熱性的假設資料檔。每一個觀察值代表對合金之一的不同檢定；記錄了讓軸承失效的溫度。
- **cereal.sav。** 這是有關對 880 人的早餐喜好進行訪談的假設資料檔，也記下他們的年齡、性別、婚姻狀況、和是否有活躍的生活型態 (根據他們是否一週運動兩次)。每一個觀察值代表一位不同的應答者。
- **clothing_defects.sav。** 這是有關一家服裝工廠品質管制過程的假設資料檔。由該工廠所生產的每一批產品中，檢查員取出一件服裝的樣本並計算不合格的服裝個數。
- **coffee.sav。** 本資料檔是關於六種冰咖啡品牌的感覺印象(Kennedy, Riquier, 和 Sharp, 1996)。對 23 種冰咖啡中每一種的印象屬性，由群眾來選取依其屬性描述的所有品牌。該六種品牌已標示為 AA、BB、CC、DD、EE、和 FF，以保持機密。
- **contacts.sav。** 這是有關一群公司電腦銷售代表聯絡清單的假設資料檔。每一個聯絡人依他們在公司所服務的部門及其公司的等級而分類。最後一次銷售的金額、到最後一次銷售的時間、和該聯絡人公司的規模也都被列入記錄。
- **creditpromo.sav。** 這是有關一家百貨公司致力於評估近期信用卡促銷活動效果的假設資料檔。為達此目標，隨機選取了 500 位持卡人。有半數收到廣告，促銷在未來三個月購買將獲得降低利率的優惠。半數收到標準的週期性廣告。
- **customer_dbase.sav。** 這是有關一家公司致力於使用其資料倉庫的資訊來對最有可能回應的客戶提供優惠的假設資料檔。隨機選取客戶庫的子集，提供優惠，再將他們的回應記錄下來。
- **customer_information.sav。** 本檔案是包含客戶郵寄資訊的假設資料檔，例如姓名和地址。
- **customer_subset.sav。** 80 個 customer_dbase.sav 的觀察值子集。
- **customers_model.sav。** 本檔案包含一市場行銷活動所鎖定之個人的假設資料。這些資料包含人口資訊、購買歷史摘要、和每一個人是否對該活動有回應。每一個觀察值代表一位不同的個人。
- **customers_new.sav。** 本檔案包含一市場行銷活動潛在候選人之個人的假設資料。這些資料包含每一位個人的人口資訊和購買歷史摘要。每一個觀察值代表一位不同的個人。
- **debate.sav。** 這是有關一項政治辯論會參與者辯論前和辯論後接受調查之成對反應的假設資料檔。每一個觀察值對應至一位不同的應答者。
- **debate_aggregate.sav。** 這是將 debate.sav 中之反應作整合的假設資料檔。每一個觀察值對應至辯論前和辯論後對偏好之交叉分類的反應。
- **demo.sav。** 這是有關提供郵寄每月優惠之購買客戶資料庫的假設資料檔。記錄了客戶是否對該優惠回應，以及各種的人口資訊。
- **demo_cs_1.sav。** 這是有關一家公司致力於匯編調查資訊資料庫之第一步的假設資料檔。每一個觀察值對應至一個不同的城市，也記錄了其地區、省、區、和城市識別。
- **demo_cs_2.sav。** 這是有關一家公司致力於匯編調查資訊資料庫之第二步的假設資料檔。每一個觀察值對應至在第一步中選取的城市中的一個不同的家庭單位，也記錄了其地區、省、區、分區、和單位識別。也納入了由該設計的前兩階段所得之取樣資訊。
- **demo_cs.sav。** 這是包含以複合取樣設計所收集之調查資訊的假設資料檔。每一個觀察值對應至一個不同的家庭單位，也記錄了各種的人口和取樣資訊。

- **dmdata.sav。** 這是包含直效行銷公司之人口和購買資訊的假設資料檔。dmdata2.sav 包含收到測試郵件的連絡人子集資訊，而 dmdata3.sav 則包含剩下未收到測試郵件的連絡人資訊。
- **dietstudy.sav。** 本假設資料檔包含對「Stillman 飲食法」(Rickman, Mitchell, Dingman, 和 Dalen, 1974) 研究的結果。每一個觀察值對應至一個不同的受試者，並記錄下他或她飲食法前、後之體重 (磅) 和三酸甘油酯水準 (毫克/100 毫升)。
- **dvdplayer.sav。** 這是有關新 DVD 播放器開發的假設資料檔。市場行銷團隊使用原型收集了焦點組別資料。每一個觀察值對應至不同調查到的使用者，並記錄下一些有關他們的人口資訊和他們對有關原型問題的回應。
- **german_credit.sav。** 本資料檔取自 (Blake 和 Merz, 1998) 艾文 (Irvine) 在加州大學機器學習資料庫儲存器的「德國信用」資料集。
- **grocery_1month.sav。** 本假設資料檔是將 grocery_coupons.sav 資料檔和每週購買的「彙總」，因此每一個觀察值對應至一個不同的客戶。結果部份每週變更的變數消失了，而目前所記錄的銷售量是在研究的四週期間銷售量之總和。
- **grocery_coupons.sav。** 這是包含某連鎖雜貨店想要知道他們客戶購買習慣所收集之調查資料的假設資料檔。每一個客戶被追蹤了四週，每一個觀察值對應至一個不同的客戶-週，並記錄有關客戶在何處及如何購物的資訊，包含那一週在雜貨店花了多少錢。
- **guttman.sav。**Bell(Bell, 1961) 以此表說明可能的社會團體。Guttman (Guttman 值, 1968) 過去曾使用此表的一部分，在這部分中有 5 個變數，分別說明 7 個理論社會團體的社會互動、團體歸屬感、成員實際接觸和關係正式性，而這 7 個群組包括：群眾 (例如，足球場上的人)、觀眾 (例如在戲院中和課堂上的人)、公眾 (例如，報紙讀者和電視觀眾)、暴民 (和群眾相似，但互動較為激烈)、原級團體 (親密性)、次級團體 (自願性) 和現代社群 (因親密的身體接近而導致鬆散的結盟和特殊服務的需求)。
- **health_funding.sav。** 這是包含醫療保健基金 (每 100 個人口的金額)、疾病率 (每 10,000 個人口的比率)、造訪醫療保健機構的比例 (每 10,000 個人口的比率) 的假設資料檔。每一個觀察值代表一個不同的城市。
- **hivassay.sav。** 這是有關一家製藥實驗室致力於開發一種偵測 HIV 感染快速檢驗的假設資料檔。檢驗結果是八個紅色加深的陰影，陰影愈深表示感染的可能性愈大。進行了一項實驗室的試驗，在 2,000 個血液樣本中，有半數遭到 HIV 的感染，而半數則未感染。
- **hourlywagedata.sav。** 這是有關在辦公室和醫院任職的護士依經驗水準不同之鐘點費的假設資料檔。
- **insurance_claims.sav.** 這是有關一家保險公司想要建立模式來標示可疑及可能的詐欺理賠之假設資料檔。每一個觀察值代表個不同的理賠。
- **insure.sav。** 這是有關一家保險公司正在研究表示客戶是否必定理賠 10 年壽險合約之風險因子的假設資料檔。在資料檔中的每一個觀察值代表二份合約，其一記錄了理賠而另一則否，二者的年齡和性別相符。
- **judges.sav。** 這是有關受過訓練的裁判 (加上一位熱心人士) 為 300 個體操表演評分的假設資料檔。每一列代表一個不同的表演；裁判們觀看相同的表演。
- **kinship_dat.sav。**Rosenberg 與 Kim (Rosenberg 和 Kim, 1975) 致力於分析 15 個親屬關係稱呼 (姑/姨、兄弟、堂/表兄弟姐妹、女兒、父親、孫女、祖父、祖母、孫子、母親、姪子/外甥、姪女/外甥女、姐妹、兒子、叔/舅父)。他們請四組大學生 (兩組女性、兩組男性) 根據其相似性來分類整理這些稱謂。他們請其中兩組 (一組

女性、一組男性）作兩次分類整理，第二次要根據與第一次不同的準則進行分類整理。因此，總共得到六個「來源」。每一個來源對應至一個 15 × 15 的相似性矩陣，其儲存格等於來源中人數減去物件在該來源中分為同組的次數。

- **kinship_ini.sav。** 本資料檔包含 kinship_dat.sav 之三維解的起始組態。
- **kinship_var.sav。** 本資料檔包含自變數「性別」、「世代」、和可用來解讀 kinship_dat.sav 解答維度的（分離）「度」。尤其，它們可用來將解答空間限制為這些變數的線性組合。
- **marketvalues.sav。** 本資料檔有關於一項在伊立諾州阿爾岡京 (Algonquin, Ill.) 的新屋開發案自 1999 年至 2000 年之房屋銷售情況。這些銷售與公共記錄有關。
- **nhis2000_subset.sav。** 「國民健康訪問調查 (NHIS)」為美國民間人口的一大型民眾調查。其以具全國代表性的家庭為樣本，面對面的完成訪問。而取得各家庭中成員的人口統計學資訊及健康行為、健康狀態方面等觀察報告。本資料檔包含一個 2000 年調查資訊的子集。國家衛生統計中心。2000 年「國民健康訪問調查 (NHIS)」。公用資料檔案和文件。ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年曾存取。
- **ozone.sav。** 本資料包含對六個氣象變數所作的 330 個觀察值，以自其餘的變數中預測臭氧濃度。先前研究人員中，(Breiman 和 Friedman 檢定(F), 1985)、(Hastie 和 Tibshirani, 1990) 在這些會阻礙標準迴歸方式的變數中發現非線性。
- **pain_medication.sav。** 本假設資料檔包含治療慢性關節炎疼痛之消炎藥物臨床試驗的結果。特別關注於藥物發生作用的時間以及它是如何與現用藥物作比較。
- **patient_los.sav。** 本假設資料檔包含對因可能為心肌梗塞 (MI，或「心臟病」）入院病患的治療記錄。每一個觀察值對應至一個不同的病患並記錄許多與其留院期間有關的變數。
- **patlos_sample.sav。** 本假設資料檔包含病患在為心肌梗塞 (MI，或「心臟病」）治療期間接受血栓溶解治療的治療記錄樣本。每一個觀察值對應至一個不同的病患並記錄許多與其留院期間有關的變數。
- **polishing.sav。** 這是取自「資料和故事圖書館」的「Nambeware 打磨時間」資料檔。它是有關一家金屬餐具製造商 (Nambe Mills, 聖塔非, 新墨西哥州）致力於規劃其生產排程。每一個觀察值代表生產線上一個不同的產品。每一個產品都記錄下直徑、打磨時間、價格、和產品類別。
- **poll_cs.sav。** 這是有關民意測驗專家致力於確定交付立法之前公眾對法案支持水準的假設資料檔。觀察值對應至登記選民。每一個觀察值記錄下選民的郡、鎮、和他居住的鄰近範圍。
- **poll_cs_sample.sav。** 本假設資料檔包含列於 poll_cs.sav 中的選民樣本。樣本是根據在 poll.csplan 計劃檔中指定的設計來取得，而本資料檔記錄了包含機率和樣本權重。不過，請注意，由於取樣計劃採用到機率 - 比例 - 大小 (PPS) 方法，也用到一個包含聯合選擇機率的檔案 (poll_jointprob.sav)。其他與選民人口及其對提議法案之意見有關的變數都在取樣後收集並加入資料檔中。
- **property_assess.sav。** 這是有關郡財產估價人員致力於對限定資源保持財產價值評估維持最新的假設資料檔。觀察值對應至郡內過去一年銷售的財產。資料檔中的每一個觀察值記錄了財產所在的鎮、上次訪查該財產的估價人員、自那次評估後經過的時間、當時定的估價、和該財產銷售價值。

- **property_assess_cs.sav。** 這是有關州財產估價人員致力於對限定資源保持財產價值評估維持最新的假設資料檔。觀察值對應至州中的財產。資料檔中的每一個觀察值記錄了郡、鎮、和財產所在的鄰近範圍、自最後一次評估後經過的時間、和當時定的估價。
- **property_assess_cs_sample.sav。** 本假設資料檔包含列於 property_assess_cs.sav 中的財產樣本。樣本是根據在 property_assess.csplan 計劃檔中指定的設計來取得，而本資料檔記錄了包含機率和樣本權重。另外的變數「目前價值」是在取樣後收集並加入資料檔中。
- **recidivism.sav。** 這是有關政府法令執行機構致力於瞭解其轄區內之再犯率的假設資料檔。每一個觀察值對應至一個先前的違法者並記錄其人口資訊、第一次犯罪的一些細節、然後是直到第二次被捕的時間（如果它發生在第一次被捕的兩年之內）。
- **recidivism_cs_sample.sav。** 這是有關政府法令執行機構致力於瞭解其轄區內之再犯率的假設資料檔。每一個觀察值對應到一個先前的違法者，在 2003 年六月第一次被捕後釋放，並記錄其人口資訊、第一次犯罪的一些細節、和第二次被捕日期（如果它發生在 2006 年六月之前）。違法者是根據在 recidivism_cs.csplan 中所指定的取樣計劃之樣本部門來選取；由於取樣計劃採用到機率 - 比例 - 大小 (PPS) 方法，也用到一個包含聯合選擇機率的檔案 (recidivism_cs_jointprob.sav)。
- **rfm_transactions.sav。** 本檔案是包含購買交易資料的假設資料檔，包括購買日期、購買項目及每一項交易的金額。
- **salesperformance.sav。** 這是有關評估兩個新售貨員訓練課程的假設資料檔。六十個員工，分成三個組別，全部接受標準訓練。此外，組別二得到技術訓練；組別三則是實務輔導簡介。每一個員工在訓練課程結束時接受測驗並記錄他們的分數。在資料檔中每一個觀察值代表一個不同的訓員，並記錄他們所分派的組別和他們在測驗中得到的分數。
- **satisf.sav。** 這是有關一家零售公司在 4 個商店位置所作之滿意度調查的假設資料檔。總共有 582 位客戶接受調查，每一個觀察值代表一位客戶的反應。
- **screws.sav。** 這個資料檔包含螺絲釘、螺栓、螺帽和圖釘之特色的資訊(Hartigan, 1975)。
- **shampoo_ph.sav。** 這是有關一家美髮產品工廠品質管制過程的假設資料檔。在固定的時間間隔，記錄下六個不同輸出批次的測量和它們的 pH 值。目標範圍是 4.5 - 5.5。
- **ships.sav。** 一個在別處 (McCullagh et al。, 1989) 出現和分析過，有關商船因風浪所造成損壞的資料集。事件次數可建立模式為以 Poisson 率發生，給定船型、建造期間、和服務期間。以因子交叉分類所形成的表格的每一個儲存格服務月數的整合，提供了暴露於風險之值。
- **site.sav。** 這是有關一家公司致力於為事業擴展選擇新地點的假設資料檔。您們僱請兩位顧問分別評估該地點，除了一份廣泛的報告之外，他們還要將每個地點摘要為前景「佳」、「可」、或「差」。
- **smokers.sav。** 本資料檔是由「1998 年全國家庭毒品濫用調查」中摘錄，且是美國家庭的機率樣本。(http://dx.doi.org/10.3886/ICPSR02934) 因此，在分析本資料檔的第一步應該是將資料加權以反映母群體傾向。
- **stroke_clean.sav。** 本假設資料檔包含一個醫療資料庫，其在以「資料準備」選項中的程序清理之後的狀態。
- **stroke_invalid.sav。** 本假設資料檔包含一個醫療資料庫的起始狀態並包含幾個資料輸入錯誤。

- **stroke_survival。** 本假設資料檔是有關缺血性中風的病患，其在結束康復計畫後存活時間方面，面臨許多挑戰。中風後，記載了心肌梗塞、缺血性中風、或出血性中風的發生，以及事件記錄的時間。由於它只包含在康復計劃所管制的中風存活的病患，此樣本的左側被截斷。
- **stroke_valid.sav。** 本假設資料檔包含一個醫療資料庫，在其值以「驗證資料」程序檢查之後的狀態。它仍包含可能的異常觀察值。
- **survey_sample.sav。** 本資料檔包含調查資料，包括人口資料和各種態度測量。雖然已修改一些資料數值，且為人口資料之目的新增了一些額外的虛構變數，但是資料仍是以「1998 NORC 基本社會調查」的變數子集為基礎。
- **telco.sav。** 這是有關一家電信公司致力於在客戶庫中減少顧客不忠的假設資料檔。每一個觀察值對應至一位不同的客戶並記錄不同的人口資料和服務使用方式資訊。
- **telco_extra.sav。** 本資料檔類似於 telco.sav 資料檔，但「任期」的對數轉換客戶花費變數已予刪除，並更換為標準的對數轉換客戶花費變數。
- **telco_missing.sav。** 本資料檔是 telco.sav 資料檔的子集，不過某些人口資料值已更換為遺漏值。
- **testmarket.sav。** 本假設資料檔有關於一家速食連鎖店計劃在菜單中加入新的項目。有三個可能的活動來促銷此新產品，所以該新項目在幾個隨機選取市場中的地點作介紹。在每一個地點使用不同的促銷，並記錄該新項目前四週的每週銷售量。每一個觀察值對應至一個不同的地點-週。
- **testmarket_1month.sav。** 本假設資料檔是將 testmarket.sav 資料檔和每週購買的「彙總」，因此每一個觀察值對應至一個不同的客戶。結果部份每週變更的變數消失了，而目前所記錄的銷售量是在研究的四週期間銷售量之總和。
- **tree_car.sav。** 這是包含人口資料和車輛購買價格資料的假設資料檔。
- **tree_credit.sav。** 這是包含人口資料和銀行放款歷史資料的假設資料檔。
- **tree_missing_data.sav** 這是包含有大量遺漏值的人口資料和銀行放款歷史資料的假設資料檔。
- **tree_score_car.sav。** 這是包含人口資料和車輛購買價格資料的假設資料檔。
- **tree_textdata.sav。** 一個只有兩個變數的簡單資料檔，主要目的在顯示變數預設狀態（在指定量測水準和數值標記之前）。
- **tv-survey.sav。** 這是有關一家電視製片廠考量是否要延長一個成功節目的播送所作之調查的假設資料檔。有 906 位應答者被問到在不同的狀況下他們是否願意觀看這個節目。每一列代表一個不同的應答者；每一行為一個不同的狀況。
- **ulcer_recurrence.sav。** 本檔案包含一項用來比較兩種防止潰瘍復發治療法功效之研究的部分資訊。它是很好的區間受限資料範例，且已在別處 (Collett, 2003) 出現和分析過。
- **ulcer_recurrence_recoded.sav。** 本檔案是將 ulcer_recurrence.sav 的資訊重新組織，以讓您為此研究的每一個區間事件機率而非只是研究目的事件機率建立模式。它已在別處 (Collett et al。, 2003) 出現和分析過。
- **verd1985.sav。** 本資料檔有關於一項調查 (Verdegaal, 1985)。在調查中記錄了來自 15 個受訪者對 8 個變數的回應。所需的變數被分成三組。集 1 包括 age 和 marital，集 2 包括 pet 和 news，集 3 包括 music 和 live。Pet 調整為多重名義量數，age 調整為次序量數，其他的變數調整為單一名義量數。

- **virus.sav。** 這是有關一家網際網路服務提供者致力於在其網路上判斷病毒之影響的假設資料檔。他們在其網路上追蹤從發現病毒直到控制威脅的這段時間，被病毒感染之電子郵件的流量（約略）百分比。
- **wheeze_steubenville.sav。** 這是空氣污染對兒童健康之影響 (Ware, Dockery, Spiro III, Speizer, 和 Ferris Jr., 1984) 縱向研究的子集。本資料包含來自俄亥俄州 Steubenville，年齡 7、8、9 和 10 歲兒童的氣喘聲狀態之重複二元測量，以及其母親在本研究的第一年是否抽煙的固定記錄。
- **workprog.sav。** 這是有關一項政府職業計劃，設法將弱勢民眾安置到較好之工作的假設資料檔。一個樣本的可能計劃參與者被追蹤，他們之中某些被選取加入本計劃，而其他的則否。每一個觀察值代表一位不同的計劃參與者。

附錄
B

Notices

Patent No. 7,023,453

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

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

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

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

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

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

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided “AS IS”, without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

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

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

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

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

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

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

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

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

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

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

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

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

參考書目

Bell, E. H. 1961. Social foundations of human behavior: (人類行為之社會基礎)Introduction to the study of sociology (社會學研究簡介). 紐約: Harper & Row.

Blake, C. L., 和 C. J. Merz. 1998. "UCI Repository of machine learning databases (機器學習資料庫 UCI 儲存器)." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., 和 J. H. Friedman 檢定(F). 1985. Estimating optimal transformations for multiple regression and correlation (估計多重迴歸與相關之最適轉換). Journal of the American Statistical Association (美國統計協會彙報), 80, .

Collett, D. 2003. Modelling survival data in medical research (模式化醫學研究中的存活資料), 2 ed. Boca Raton: Chapman & Hall/CRC.

Green, P. E., 和 V. Rao. 1972. Applied multidimensional scaling (應用多元尺度方法). Hinsdale, Ill.: Dryden Press (Dryden 出版社).

Green, P. E., 和 Y. Wind. 1973. Multiattribute decisions in marketing: (市場行銷之多重屬性決策)A measurement approach (測量方法). Hinsdale, Ill.: Dryden Press (Dryden 出版社).

Guttman 值, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points (尋找組態點最小座標空間之一般非計量技巧). Psychometrika (心理學計量報導), 33, .

Hartigan, J. A. 1975. Clustering algorithms (集群演算法). 紐約: John Wiley and Sons.

Hastie, T., 和 R. Tibshirani. 1990. Generalized additive models (概化附加模式). 倫敦: Chapman and Hall.

Kennedy, R., C. Riquier, 和 B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research (市場研究類別資料之對應分析實際應用). Journal of Targeting, Measurement, and Analysis for Marketing (市場行銷之目標訂定、測量與分析雜誌), 5, .

McCullagh, P., 和 J. A. Nelder. 1989. 概化線性模式, 第二版 ed. 倫敦: Chapman & Hall.

Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior (行為適切性與情境限制作為社會行為維度). Journal of Personality and Social Psychology (人格與社會心理學雜誌), 30, .

Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet (實行 Stillman 飲食法期間血膽固醇改變情形). Journal of the American Medical Association (美國醫學協會彙報), 228, .

Rosenberg, S., 和 M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research (多變量研究中作為資料收集程序之排序方法). Multivariate Behavioral Research (多變量行為研究), 10, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, 和 H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: (依經驗法將青少年飲食異常次組別化)A longitudinal perspective (縱向觀點). British Journal of Psychiatry (英國心理學雜誌), 170, .

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch) (更多性質資料的集合分析 (荷蘭文)). Leiden (萊頓): Department of Data Theory, University of Leiden (萊頓大學資料理論系).

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, 和 B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities (六個城市中的兒童二手煙、氣體煮食與呼吸道健康). American Review of Respiratory Diseases (美國呼吸道疾病評論), 129, .

索引