IBM SPSS Decision Trees 20

注意：使用此資訊和支援的產品之前，請先閱讀 注意事項第 99 頁 包含的一般資訊。

若新版本未聲明，則此版本便適用於 IBM® SPSS® Statistics 20 以及之後發行的所有版本和修正。

Adobe 產品的擷取畫面已取得 Adobe Systems Incorporated 之翻印許可。

Microsoft 產品的擷取畫面已取得 Microsoft Corporation 之翻印許可。

授權內容：IBM 資產

Copyright IBM Corporation 1989, 2011.

美國政府使用者有限權利：使用、複製或披露內容皆受 IBM Corp 簽署之 GSA ADP Schedule Contract 限制約束。

序

IBM® SPSS® Statistics為分析資料的強大系統。Decision Trees 的選用性附加模組能提供其他本手冊所說明的分析技術。Decision Trees 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 IBM Business Analytics

IBM Business Analytics 軟體提供完整、一致且確實的資訊，決策者可信任此資訊，並藉以改善營運績效。包括商業智慧、預測分析、財務績效和策略管理，以及分析應用程式的整合型產品組合，為目前績效提供了清晰、即時且具行動性的前瞻眼界，以及預測未來成果的能力。結合了豐富的業界解決方案、有效實證和專業服務，每種規模的組織都能引爆最高效能，確實自動化執行決策，並且交付更棒的成果。

在這項產品組合中，IBM SPSS Predictive Analytics 軟體有助於組織預測未來事件，並且針對前瞻概念提前行動，創造更棒的營運成果。全球的商業、政府和學術客戶相當倚重 IBM SPSS 技術所帶來的競爭優勢，藉此做為吸引、保有和發展更多客戶，同時降低可能的不實詐欺風險。藉由將 IBM SPSS 軟體併入每天作業，這些組織成為預測型企業 – 足以駕馭決策並使決策自動化處理，以符合營運目標，並且達到可測知的競爭優勢。如需更多資訊，或是聯絡代表人員，請造訪 http://www.ibm.com/spss。

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 IBM Corp. 產品在使用上的協助，或是支援硬體環境的安裝說明。若要取得技術支援，請參閱 IBM Corp. 網站內容，網址：http://www.ibm.com/support。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。

針對學生用戶的技術支援

如果您是使用任何 IBM SPSS 軟體產品之學生版、學術版或研究套件版本的學生，請參閱適用於學生的特殊線上「教育解決方案 (http://www.ibm.com/spss/rd/students/)」頁面。如果您是使用 IBM SPSS 軟體之大學提供副本的學生，請聯絡您大學的 IBM SPSS 產品協調人員。

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。請備妥您的序號以供識別。

訓練研討會

IBM Corp. 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。如需研討會的詳細資訊，請移至 http://www.ibm.com/software/analytics/spss/training。

Copyright IBM Corporation 1989, 2011.

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

部 I: 使用手冊

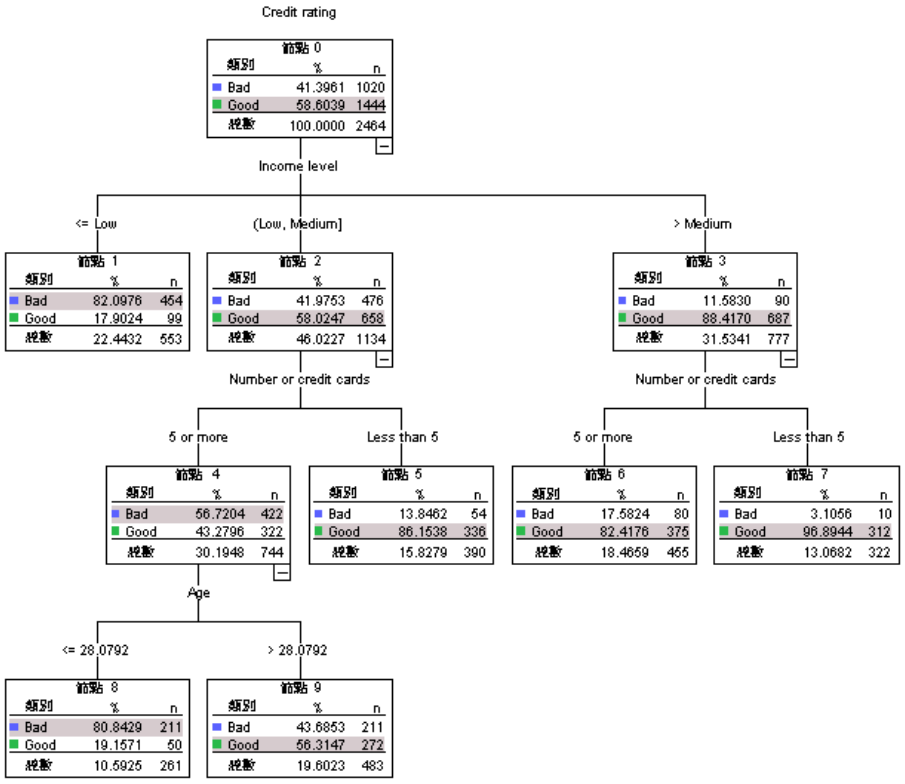
部 II: 範例

部 I:
使用手冊

章節
1

建立決策樹狀結構

圖表 1-1
決策樹狀結構



「決策樹狀結構」程序會建立樹狀結構的分類模式。它會根據自（預測值）變數的值，將觀察值分成組別，或依變數（目標）的預測值。這個程序會提供用於解釋與確認分類分析的驗證工具。

這個程序可以用於：

分段。 識別有可能是特殊群組成員的人員。

分層。 將觀察值指定給其中一個類別，例如高風險、中風險與低風險群組。

預測。 建立規則並用這些規則來預測未來的事件，例如某人可能會借貸，或是汽車或房屋的潛在重新銷售值。

資料縮減與變數篩檢。 從一個大型的變數集合中選取一個有用的預測值子集，用於建立一個正式的參數模式。

交互作用識別。 識別只與特定次組別有關的關係並在正式的參數模式中指定這些項目。

Copyright IBM Corporation 1989, 2011.

類別合併與離散化連續的變數。 使用遺失最少資訊的方式將組別預測值類別與連續變數重新編碼。

範例。 假設某家銀行打算根據信用申請人是否有合理的信用風險，來將這些申請人加以分類。根據各種因素，包括過去客戶的已知信用評等，您就可以建立一個模式來預測未來的客戶是否有可能進行借貸。

樹狀結構分析會提供一些引人注意的功能：

- 它可以讓您識別具有高風險或低風險的同質組別。
- 它可以更容易建構有關個別觀察值進行預測的規則。

資料考量

資料。 依變數和自變數可以是：

- **名義。** 當變數數值代表實質上並未等級化的類別時（例如，有員工工作的公司部門），則此變數可視為名義。名義變數的範例包括地區、郵遞區號以及宗教團體。
- **次序。** 當變數數值代表實質上已等級化的類別時（例如，服務滿意度從非常不滿意到非常滿意分級），則此變數可視為次序。次序變數的範例包括代表滿意度或信賴程度的態度分數、以及偏好等級分數。
- **尺度。** 若一變數可視為尺度（連續），表示它的的數值代表含有實際意義矩陣的已排列順序類別，因此適合比較數值之間的距離。尺度變數的範例包括以年份表示的年齡和以千元為單位的收入。

頻率加權 如果加權生效的話，則分數加權就會捨入為最接近的整數，所以，加權值少於 0.5 的觀察值就會被指定一個 0 的加權，進而從分析中被排除在外。

假設。 這個程序會假設已經將適當的測量水準指定給所有分析變數，而且某些功能會假設包含在分析中的依變數的所有值都已經定義數值標記。

- **測量水準。** 測量水準會影響樹狀結構計算作業，因此所有變數都應該指定適當的測量水準。根據預設，數值變數是假設為尺度變數而字串變數則假設為名義變數，它們可能無法精確反映真正測量的水準。變數清單中各變數旁圖示會指明變數類型。


尺度


名義


次序

您可以藉由在來源變數清單按一下滑鼠右鍵，從內容功能表選取測量水準，暫時變更變數的測量水準。

- **數值標記。** 這個程序的對話方塊會假設類別（名義、次序）依變數的所有非遺漏值都已經定義數值標記，或是都沒有定義數值標記。除非類別依變數中至少有兩個非遺漏值具有數值標記，否則某些功能將無法使用。如果至少兩個非遺漏值已定義數值標記，當有任何觀察值具有其他無數值標記的數值時，該觀察值會從分析中排除。

若要取得決策樹狀結構

▶ 從功能表選擇：
分析(A) > 分類 > 樹...

圖表 1-2
「決策樹狀結構」對話方塊

▶ 選取依變數。

▶ 選取一個或多個自變數。

▶ 選取一個成長方法。

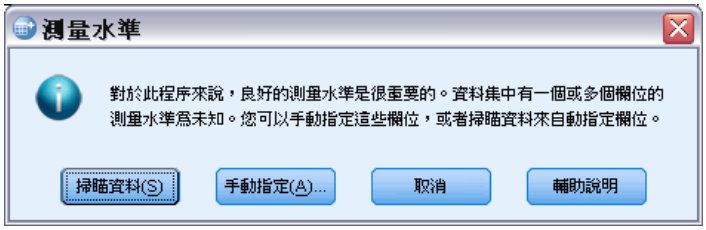
您可以：

- 變更來源清單中任何變數的測量水準。
- 強制自變數清單中的第一個變數進入模式中當做第一個分割變數。
- 選取定義觀察值影響樹狀結構成長過程之程度的影響變數。觀察值的影響變數數值較低，則影響力較小；反之則影響力較大。影響變數值必須為正數。
- 驗證樹狀結構。
- 自訂樹狀結構成長條件。
- 將終端節點數、預測值以及預測機率另存成變數。
- 以 XML (PMML) 格式儲存模式。

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 1-3
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。

變更測量水準

▶ 在來源清單的變數上按一下滑鼠右鍵。

▶ 選取快顯功能表上的測量水準。

這會暫時變更測量水準，以供在「決策樹狀結構」程序中使用。

成長方法。

可用的成長方法包括：

CHAID。 卡方自動交互作用偵測。CHAID 會在每個步驟中，選擇與依變數具有最強交互作用的自（預測）變數。若與相關的依變數沒有明顯不同，則會合併每個預測變數的類別。

Exhaustive CHAID。 一種 CHAID 的修正，其會檢驗每個預測值的所有可能分割。

CRT。 分類與迴歸樹狀結構。CRT 會盡量將資料分割成與依變數相關的同質資料片段。在終端節點中，所有觀察值皆具有相同的依變數值，因此會是「純」同質節點。

QUEST。 快速、不偏且有效的統計之樹狀結構。此方法不但計算快速，而且能避免如其他方法偏好有許多類別的預測變數。只有在名義依變數才能指定 QUEST。

每一個方法都有其優點與限制，包括：

	CHAID*	CRT	QUEST
以卡方分配為基礎**	C		
代理自（預測值）變數		C	C

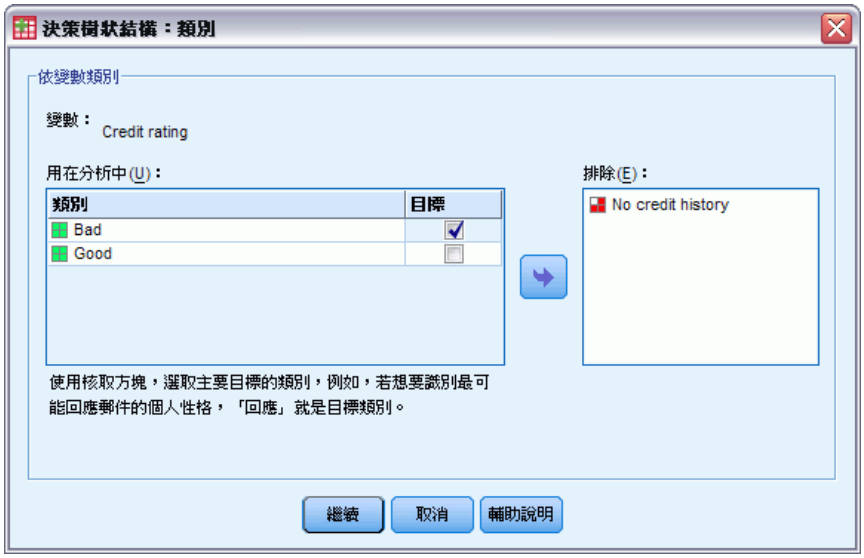
	CHAID*	CRT	QUEST
樹狀結構修正		C	C
多因子節點分割	C		
二元節點分割		C	C
影響變數	C	C	
事前機率		C	C
錯誤分類成本	C	C	C
快速計算	C		C

*包括 Exhaustive CHAID。

**QUEST 也會將卡方量數用於名義自變數。

選取類別

圖表 1-4
「類別」對話方塊



如果是類別（名義、次序）依變數，您可以：

- 控制要包含在分析中的類別。
- 識別相關的目標類別。

包含/排除類別

您可以將分析限制為依變數的特定類別。

- 在「排除」清單中之依變數值的觀察值不會包含在分析中。
- 如果是名義依變數，您也可以在分析中包含使用者遺漏的類別(依照預設值，使用者遺漏的類別會顯示在「排除」清單中)。

目標類別

已選取的（已核取的）類別會被視為分析中主要相關的類別。例如，如果您主要是要識別最有可能借貸的個人，您可以選取信用評等類別「差」來當做目標類別。

- 沒有預設的目標類別。如果沒有選取任何類別，某些分類規則選項與獲利相關選項就無法使用。
- 如果選取多個類別，就會為每一個目標類別產生個別的獲利表與圖表。
- 將一個或多個類別指定為目標類別，對於樹狀模式、風險估計或錯誤分類結果並不會有任何影響。

類別及數值標記

這個對話方塊需要依變數的已定義數值標記。除非類別依變數的至少兩個值已經定義數值標記，否則無法使用這個對話方塊。

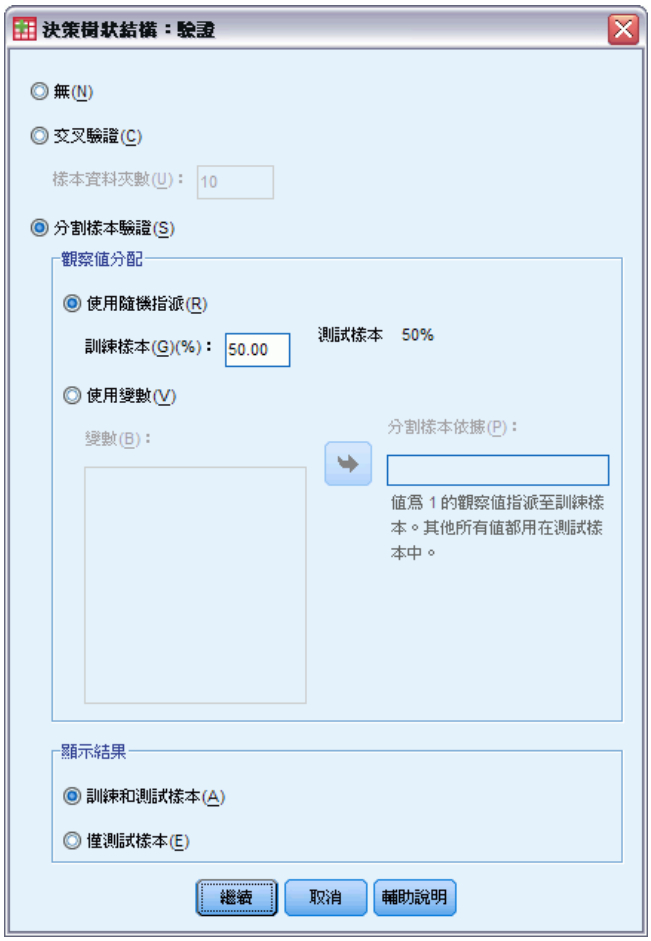
若要包含/排除類別並選取目標類別

▶ 在主要的「決策樹狀結構」對話方塊中，選取具有兩個或多個數值標記的類別（名義、次序）依變數。

▶ 按一下「類別」。

驗證 (V)

圖表 1-5
驗證對話方塊



驗證可以讓您評估樹狀結構對大型母群概化的程度。有兩種驗證方法可供使用：交叉驗證與分割樣本驗證。

交叉驗證 (C)

交叉驗證會將樣本分成次樣本數或**折疊**。接著會產生樹狀結構模式，然後從每個次樣本中排除資料。第一個樹狀結構是以第一個樣本折疊中以外的所有觀察值為基礎，第二個樹狀結構是以第二個範例折疊中以外的所有觀察值為基礎，依此類推。對於每一個樹狀結構，都會藉由將樹狀結構套用至在產生時被排除的次樣本來評估錯誤分類風險。

- 您最多可以指定 25 個樣本折疊。如果值越高，則從每一個樹狀結構模式中排除的觀察值數量也就越少。
- 交叉驗證會產生一個單一、最終的樹狀結構模式。最終樹狀結構的交叉驗證風險評估是計算所有樹狀結構風險的平均值。

分割樣本驗證

透過分割樣本驗證，就可以使用訓練樣本來產生模式，以及在保留樣本上測試模式。

- 您可以指定一個訓練樣本大小（以總樣本大小的百分比表示），以及指定一個會將樣本分割成訓練與測試樣本的變數。
- 如果您使用變數來定義訓練與測試樣本，則具有變數值 1 的觀察值就會被指定給訓練樣本，而所有其他觀察值則會被指定給測試樣本。變數不可以是依變數、加權變數、影響變數或強制自變數。
- 您可以同時顯示訓練樣本與測試樣本，或是僅顯示測試樣本。
- 在小型資料檔案（含有少量觀察值的資料檔案）上使用分割樣本驗證時，必須小心。小型的訓練樣本大小可能會產生品質不佳的模式，因為這些樣本大小在某些類別中可能沒有足夠的觀察值，所以無法適當地讓樹狀結構成長。

樹狀結構成長條件

可用的成長條件是根據成長方法、依變數的測量水準或兩者的組合而定。

成長限制

圖表 1-6
「條件」對話方塊，「成長限制」索引標籤

「成長限制」索引標籤可以讓您限制樹狀結構中的水準數量，以及控制個父節點與子節點的最小觀察值數量。

最大樹狀結構深度。 控制根節點底下成長的最大水準數量。「**自動**」設定會將樹狀結構限制在 CHAID 與 Exhaustive CHAID 方法之根節點底下的三個水準數量，以及 CRT 與 QUEST 方法的五個水準數量。

最小觀察值個數。 控制節點的最小觀察值個數。沒有符合這些條件的節點將不會被分割。

- 如果增加最小值的數字，將會產生具有較少節點的樹狀結構。
- 如果減少最小值的數字，將會產生具有較多節點的樹狀結構。

對於具有少數觀察值的資料檔案，父節點的預設觀察值 100 與子節點的預設觀察值 50 有時候可能會在根節點底下產生不具有任何節點的樹狀結構；在這種情況下，如果減少最小值的數字，可能會產生更多有用的結果。

CHAID 條件

圖表 1-7
「條件」對話方塊，CHAD 索引標籤

對於 CHAID 與 Exhaustive CHAID 方法，您可以控制：

顯著性水準。您可以控制用於分割節點與合併類別的顯著值。對這兩個條件而言，預設的顯著性水準都是 0.05。

- 為了要分割節點，值就必須大於 0 並小於 1。較小的值會產生節點較少的樹狀結構。
- 為了要合併類別，則值必須大於 0 並小於或等於 1。若要避免合併類別，請指定一個 1 的值。如果是尺度自變數，這是表示最終樹狀結構中變數類別的數量就是區間的指定數量（預設值為 10）。 如需詳細資訊，請參閱第 10 頁CHAID 分析的尺度區間。

卡方統計量。 如果是次序依變數，則用於決定節點分割與類別合併的卡方，就是使用概似比方法所計算的。如果是次序依變數，您可以選擇方法：

- **Pearson。** 這個方法會提供更快速的計算，但是用在小型樣本時則必須小心。此為預設的方法。
- **概似比。** 這個方法比 Pearson 方法更為穩健，但是在計算時必須花費較多的時間。對於小型的樣本，這是較佳的方法。

模式估計。 對於名義與次序依變數，您可以指定：

- **最大疊代次數。** 預設值為 100。如果樹狀結構因為已經到達最大疊代次數而停止成長，您可以增加最大值的數字，或是變更一個或兩個控制樹狀結構成長的條件。
- **儲存格期望次數中的最少變更。** 值必須大於 0 並且小於 1。預設值為 0.05。較低的值會產生具有較少節點的樹狀結構。

使用 Bonferroni 方法調整顯著值。 對於多重比較而言，用於合併與分割條件的顯著值是透過使用 Bonferroni 方法所調整。此為預設值。

允許在節點中重新分割已合併的類別。除非您明確地防止類別進行合併，否則程序會嘗試一起合併自（預測值）變數類別以產生會描述模式之最簡單的樹狀結構。這個選項允許程序重新分割已經合併的類別（如果那樣會提供更佳解答的話）。

CHAID 分析的尺度區間

圖表 1-8
「條件」對話方塊，「區間」索引標籤

在 CHAID 分析中，在進行分析之前，尺度自（預測值）變數一定會先被分成離散的組別（例如 0 - 10、11 - 20、21 - 30 等）。您可以控制群組的初始/最大數（即使程序可能會在初始分割之後合併連續的組別）：

- **固定數。** 所有的尺度自變數一開始都會被分成相同的組別個數。預設值是 10。
- **自訂。** 每一個尺度自變數一開始都會被分成為該變數指定的組別數。

若要指定尺度自變數的區間

▶ 在主要的「決策樹狀結構」對話方塊中，選取一個或多個尺度自變數。

▶ 如果是成長方法，請選取「CHAID」或「Exhaustive CHAID」。

- ▶ 按一下「條件」。
- ▶ 按一下「區間」索引標籤。

在 CRT 與 QUEST 分析中，所有的分割都是二元分割，且尺度與次序自變數都是以相同方式處理，所以，您無法為尺度自變數指定區間數。

CRT 條件

圖表 1-9
「條件」對話方塊，CRT 索引標籤

CRT 成長方法會嘗試最大化節點內的同質性。如果某個節點無法表示觀察值的同質子集，就表示有**雜質**。例如，如果某個終端節點中的所有觀察值都具有依變數的相同值，由於該觀察值為純同質節點，因此不需要進一步分割。

您可以選取用來測量雜質的方法，以及進行分割節點時所需的最小雜質減少量

雜質測量。 如果是依變數，會使用雜質的最小平方差 (LSD) 測量，它會計算節點內變異數，並為任何頻率加權或影響值調整。

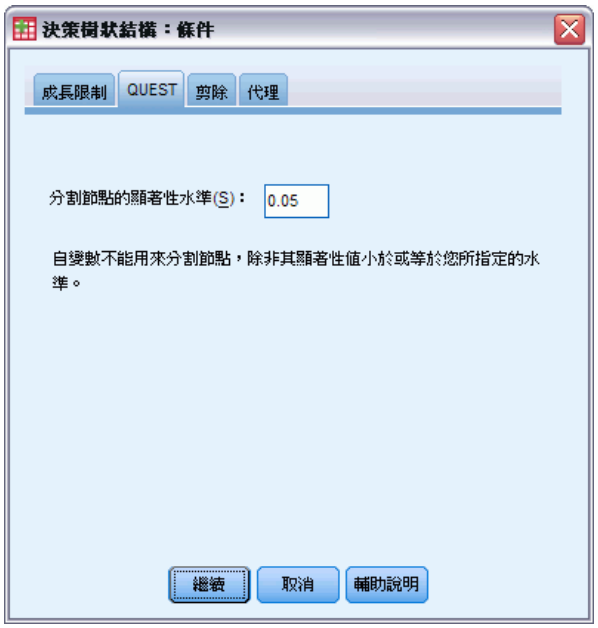
如果是類別（名義、次序）依變數，您可以選取雜質測量：

- **Gini。** 會找到與依變數值相關之子節點同質性最大化的分割。Gini 是根據依變數之每個類別成員的平方機率為基礎。當節點中的所有觀察值都落在單一個類別中時，它就會達到最小值（零）。此為預設的測量。
- **Twoing。** 依變數的類別會被分成兩個次類別組別。會找到能夠以最佳方式分開兩個組別的分割。
- **Ordered Twoing。** 與 Twoing 類似，差別在於只有相鄰類別才可以加分組。這個測量只能用於次序依變數。

改善中的最小變更。 這是分割節點時所需的最小雜質減少量。預設值是 0.0001。較高的值會產生具有較少節點的樹狀結構。

QUEST 條件

圖表 1-10
「條件」對話方塊，QUEST 索引標籤



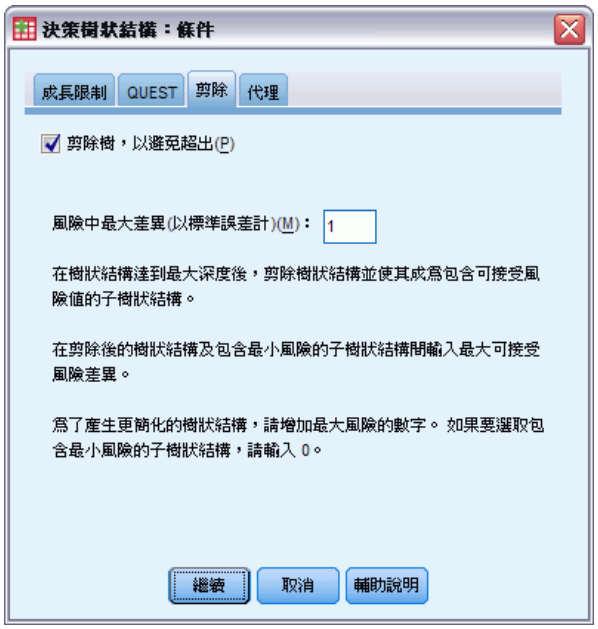
如果是 QUEST 方法，您可以指定用於分割節點的顯著性水準。除非顯著性水準小於或等於指定的水準，否則您無法使用自變數來分割節點。值必須大於 0 並且小於 1。預設值為 0.05。較小的值會從最終模式中排除更多的自變數。

若要指定 QUEST 條件

- 在主「決策樹狀結構」對話方塊中，選取名義依變數。
- 如果是成長方法，請選取「QUEST」。
- 按一下「條件」。
- 按一下「QUEST」索引標籤。

修正樹狀結構

圖表 1-11
「條件」對話方塊，「修正」索引標籤



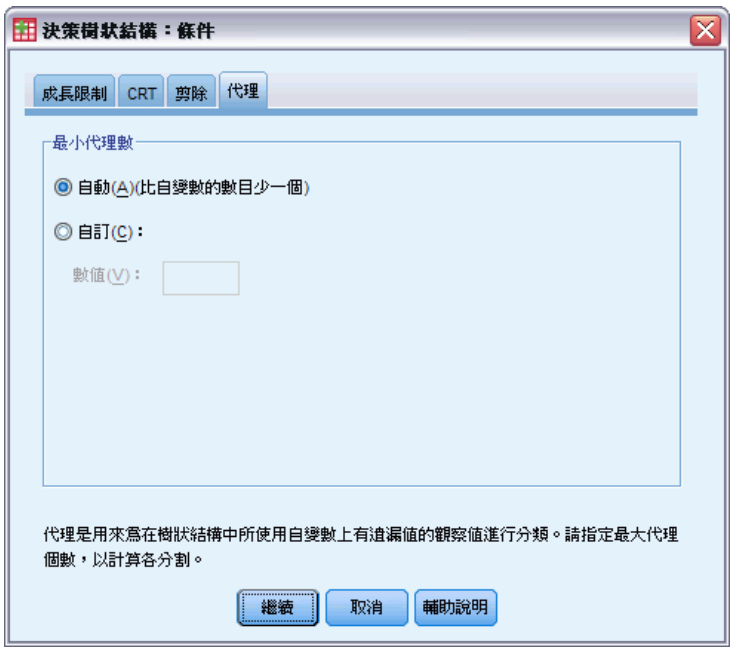
透過 CRT 與 QUEST 方法，您可以藉由**修正**樹狀結構來避免模型過適的情況：在達到停止條件時，樹狀結構會就會停止成長，然後樹狀結構會根據風險中所指定的最大差異自動修正為最小的子樹狀結構。風險值是以標準誤來表示。預設值為 1，而且必須是正數。若要取得具有最低風險的子樹狀結構，請指定 0。

修正與隱藏節點的比較

當您建立已經修正的樹狀結構時，從樹狀結構中所修正的任何節點都無法用在最終樹狀結構中。您可以使用互動方式來隱藏或顯示在最終樹狀結構中所選擇的子節點，但是您無法顯示在樹狀結構建立過程中所修正的節點。 如需詳細資訊，請參閱第 35 頁第 2 章中的樹狀編輯器。

代理

圖表 1-12
「條件」對話方塊，「代理」索引標籤



CRT 與 QUEST 可以將**代理**用於自（預測值）變數。對於該變數值已遺漏的觀察值而言，會使用其他具有與原始變數高度關聯的自變數來進行分類。這些替代的預測值稱為代理。您可以指定要用在模式中的最大代理數。

- 根據預設，最大的代理數是自變數的數量減去1 的數字。換句話說，對於每一個自變數，所有其他的自變數都可以當做代理來使用。
- 如果您不希望模式使用代理，請為代理數指定 0。

選項

可以使用的選項會因為成長方法、依變數的測量水準，及/或是否有依變數之值的已定義數值標記等而有所不同。

錯誤分類成本

圖表 1-13
「選項」對話方塊，「錯誤分類成本」索引標籤

如果是類別（名義、次序）依變數，則錯誤分類成本可以讓您包含與不正確分類相關的相對懲罰資訊。例如：

- 拒絕信用良好之客戶所花費的成本，可能不同於擴展日後會借貸之客戶的信用所花費的成本。
- 將具有高度心臟疾病風險的人員錯誤分類為具有低度心臟疾病風險人員所付出的成本，可能會高於將具有低度心臟疾病風險的人員錯誤分類為具有高度心臟疾病風險人員所付出的成本還要高。
- 將大量郵件傳送給不太可能回應的人，成本可能比較低，但是如果不將大量郵件傳送給可能會回應的人，成本相對上可能會比較高（以損失的收益而言）。

錯誤分類成本與數值標記

除非類別依變數至少有兩個值已經定義數值標記，否則無法使用這個對話方塊。

若要指定錯誤分類成本

▶ 在主要的「決策樹狀結構」對話方塊中，選取具有兩個或多個數值標記的類別（名義、次序）依變數。

▶ 按一下「**選項**」。

▶ 按一下「**錯誤分類成本**」索引標籤。

▶ 按一下「**自訂**」。

- 在網格中輸入一個或多個錯誤分類成本。輸入的值必須是非負數。(正確的分類會顯示在對角線上，而且一定是 0)。

填滿矩陣。 在許多情況中，您可能會想要讓成本變成對稱—也就是說，將 A 錯誤分類為 B 所用的成本，與將 B 錯誤分類為 A 所用的成本是一樣的。以下的控制可以讓您更輕易地指定對稱的成本矩陣：

- **複製下三角形。** 將下三角形矩陣(對角線底下）的值複製到對應的上三角形儲存格中。
- **複製上三角形。** 將矩陣下三角形（對角線底下）內的值複製到對應的上三角形儲存格中。
- **使用平均儲存格值。** 對於每一半矩陣的每一個儲存格而言，兩個值（上與下三角形）是相加之後的平均值，而這個平均值會取代原來的兩個值。例如，如果將 A 錯誤分類為 B 所付出的成本為 1，而將 B 錯誤分類為 A 所付出的成本為 3，則這個控制會使用平均值 2 ((1+3)/2 = 2) 來取代原來的兩個值。

利潤

圖表 1-14
選項對話方塊，利潤索引標籤

如果是類別依變數，您可以指定依變數水準的收益與支出值。

- 利潤是以收益減去支出的方式來計算。
- 利潤值會影響獲利表中利潤與 ROI（投資報酬率）的值，但是不會影響基本樹狀結構的模式結構。
- 收益與支出值都必須是數值，而且也都必須指定給網格中所顯示之依變數的所有類別。

利潤與數值標記

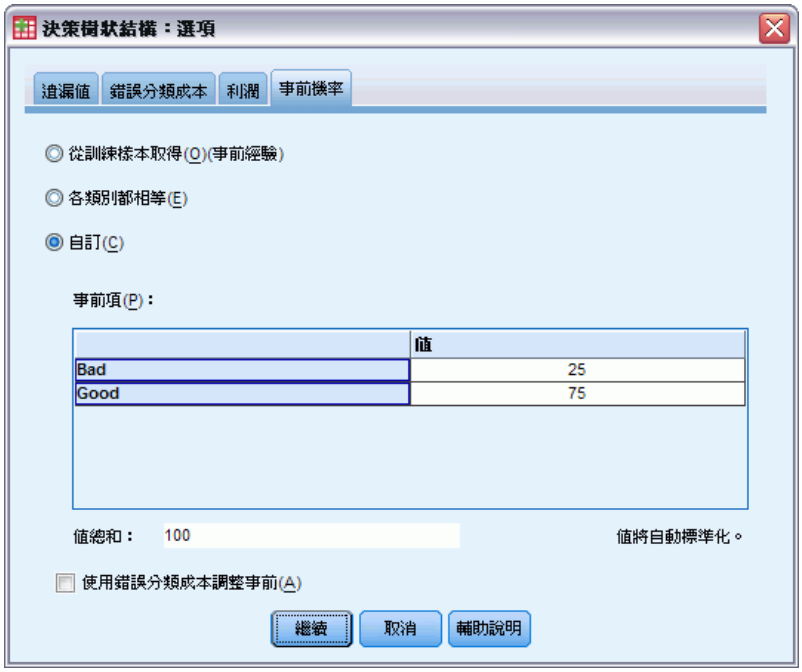
這個對話方塊需要依變數的已定義數值標記。除非類別依變數的至少兩個值已經定義數值標記，否則無法使用這個對話方塊。

若要指定利潤

- 在主要的「決策樹狀結構」對話方塊中，選取具有兩個或多個數值標記的類別（名義、次序）依變數。
- 按一下「選項」。
- 按一下「利潤」索引標籤。
- 按一下「自訂」。
- 為網格中所列出的所有依變數類別輸入收益與支出值。

事前機率

圖表 1-15
選項對話方塊，事前機率索引標籤



如果是具有類別依變數的 CRT 與 QUEST 樹狀結構，您可以指定組別成員的事前機率。**事前機率** 就是在瞭解自（預測值）變數之前，對依變數之每一個類別總體相對次數的估計。使用事前機率可以協助更正由非整體母群之取樣中的資料所造成的任何樹狀結構成長情況。

從訓練範例（經驗先驗）取得。 如果資料檔案中的依變數值分配是表示母群分配，請使用這個設定。如果您是使用分割樣本驗證，就會使用訓練樣本中的觀察值分配。

注意：由於觀察值是隨機指定給分割樣本驗證中的訓練樣本，因此無法事先知道訓練樣本中實際的觀察值分配。 如需詳細資訊，請參閱第 7 頁驗證(V)。

在所有類別保持相等。 如果依變數的類別在母群中都是顯示為相等，請使用這個設定。例如，如果一共有四個類別，則每各類別中都會有大約 25% 的觀察值。

自訂。 為網格中所列的每一個依變數類別輸入一個非負數值。值可以是比例、百分比、次數個數，或是在所有類別中表示數值分布的其他值。

使用錯誤分類成本調整先驗。 如果您定義自訂錯誤分類成本，就可以根據這些成本來調整事前機率。 如需詳細資訊，請參閱第 15 頁錯誤分類成本。

利潤與數值標記

這個對話方塊需要依變數的已定義數值標記。除非類別依變數的至少兩個值已經定義數值標記，否則無法使用這個對話方塊。

若要指定事前機率

- 在主要的「決策樹狀結構」對話方塊中，選取具有兩個或多個數值標記的類別（名義、次序）依變數。
- 如果是成長方法，請選取「CRT」或「QUEST」。
- 按一下「選項」。
- 按一下「事前機率」索引標籤。

分數

圖表 1-16
選項對話方塊，分數索引標籤

如果是具有次序依變數的 CHAID 與 Exhaustive CHAID，您可以自訂依變數之每一個類別的分數。分數會定義依變數各類別之間的順序與距離。您可以使用分數來增加或減少次序值之間的相對距離，或是變更值的順序。

- **為每個類別使用次序等級。** 依變數的最低類別會被指定一個 1 的分數，下一個較高的類別會被指定一個 2 的分數，依此類推。此為預設值。
- **自訂。** 為網格中所列的每一個依變數類別輸入一個數值分數。

範例

數值註解	原始值	分數
非技術人員	1	1
技術人員	2	4
事務人員	3	4.5
Professional	4	7
管理人員	5	6

- 分數會增加非技術人員與技術人員之間的相對距離，而且會減少技術人員與 事務人員之間的相對距離。
- 分數會將管理人員與專業人員的順序反轉。

分數與數值標記

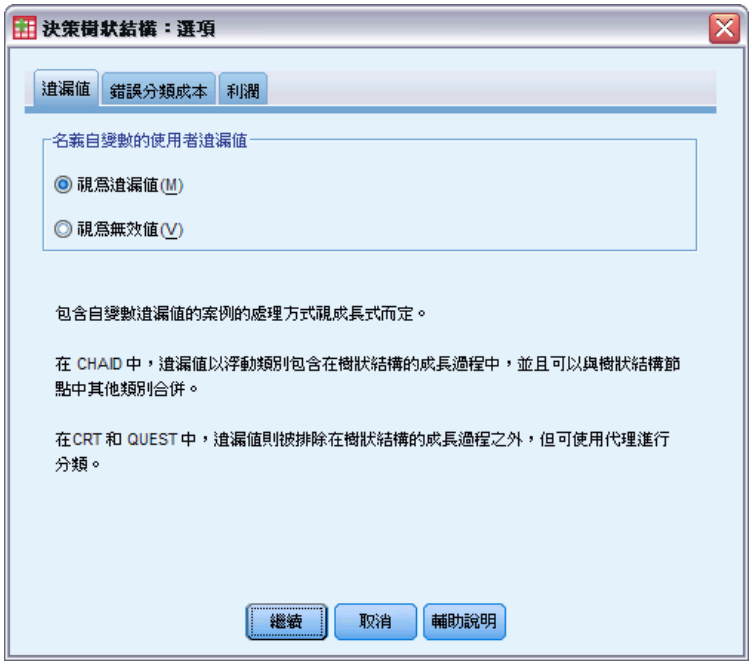
這個對話方塊需要依變數的已定義數值標記。除非類別依變數的至少兩個值已經定義數值標記，否則無法使用這個對話方塊。

若要指定分數

- ▶ 在主要的「決策樹狀結構」對話方塊中，選取具有兩個或多個已定義之數值標記的次序依變數。
- ▶ 如果是成長方法，請選取「CHAID」或「Exhaustive CHAID」。
- ▶ 按一下「選項」。
- ▶ 按一下「分數」索引標籤。

遺漏值

圖表 1-17
選項對話方塊，遺漏值索引標籤



「遺漏值」索引標籤會控制名義值、使用者遺漏值與自（預測值）變數值的處理方式。

- 次序與尺度使用者遺漏的自變數值的處理方式會因為成長方法而有所不同。
- 名義依變數的處理方式是在「類別」對話方塊中所指定。 如需詳細資訊，請參閱第 5 頁選取類別。
- 如果是次序與尺度依變數，一定會排除具有系統遺漏或使用者遺漏的依變數值的觀察值。

視為遺漏值處理。 使用者遺漏值會被視為系統遺漏值來處理。系統遺漏值的處理方式會因為成長方法而有所不同。

視為有效值處理。 名義自變數的使用者遺漏值會被視為樹狀結構成長與分類中的普通值來處理。

方法相依規則

如果某些（非全部）自變數值是系統遺漏值或使用者遺漏值：

- 如果是 CHAID 與 Exhaustive CHAID，系統遺漏與使用者遺漏的自變數值會以單一、組合的類別包含在分析中。如果是尺度與次序自變數，則演算法會先使用有效的值來產生類別，然後決定是否要將遺失的類別和其最相似的(有效的) 類別加以合併，或是將其維持為一個個別的類別。
- 如果是 CRT 與 QUEST，則具有遺失自變數值的觀察值會從樹狀結構成長過程中被排除，但是會使用代理來加以分類（如果方法中含有代理的話）。如果名義使用者遺漏值是被視為遺失值來處理的話，也會使用這個方法來處理這些值。 如需詳細資訊，請參閱第 14 頁代理。

若要指定名義、自變數使用者遺失處理

1. 在主要的「決策樹狀結構」對話方塊中，選取至少一個名義自變數。
2. 按一下「選項」。
3. 按一下「遺漏值」索引標籤。

儲存模式資訊

圖表 1-18
「儲存」對話方塊

您可以將模式的資訊儲存為工作資料檔案中的變數，也可以將整個方法以 XML (PMML) 格式儲存至某個外部檔案。

已儲存變數

終端節點數。 每個觀察值指定的終端節點。值就是樹狀結構節點數。

預測的值。 由模式所預測之依變數的類別（群組）或值。

預測的機率。 與模式的預測相關的機率。系統會為依變數的每一個類別儲存一個變數。不適用於尺度依變數。

樣本指定（訓練/測試）。 如果是分割樣本驗證，這個變數會指出訓練或測試樣本中是否有使用觀察值。訓練樣本的值為 1，測試樣本的值則為 0。除非您已經選取分割樣本驗證，否則無法使用。 如需詳細資訊，請參閱第 7 頁驗證(V)。

以 XML 格式匯出樹模式

您可以將整個樹狀結構模式儲存為 XML (PMML) 格式。您可以使用這個模式檔案，將模式資訊套用到其他資料檔案中以進行評分工作。

訓練樣本。 將模式寫入至指定的檔案。如果是分割樣本驗證樹狀結構，這是用於訓練樣本的模式。

測試樣本。 將測試樣本的模式寫入至指定的檔案。除非您已經選取分割樣本驗證，否則無法使用。

輸出

可用的輸出選項依成長方法、依變數的測量水準與其他設定而定。

樹狀結構顯示

圖表 1-19
「輸出」對話方塊，「樹狀結構」索引標籤

您可以控制樹狀結構的初始外觀，或完全隱藏樹狀結構顯示。

樹狀結構。 依照預設值，樹狀結構表是包含在「瀏覽器」中所顯示的輸出中。取消選取（取消核取）這個選項，就可以從輸出中排除樹狀結構圖。

顯示。 這些選項會控制「瀏覽器」中樹狀結構圖的初始外觀。所有這些屬性都可以藉由編輯產生的樹狀結構來加以修改。

- **方向。** 樹狀結構可以顯示為根節點在頂端，可以從上到下展開、或根節點在左右兩側，可以從左到右或從右到左展開。
- **節點內容。** 節點可以顯示表格、圖表，或同時顯示兩者。如果是類別依變數，表格會顯示次數個數與百分比，而圖表則為長條圖。如果是尺度依變數，表格會顯示平均數、標準差、觀察值數與預測的值，而圖表則為直方圖。
- **尺度。** 依照預設值，大型的樹狀結構都會自動調整，嘗試讓樹狀結構能夠符合頁面的大小。您可以指定高達 200% 的自訂尺度百分比。

- **自變數統計量。** 如果是 CHAID 與 Exhaustive CHAID，統計量包括 F 值（用於尺度依變數）或卡方值（用於類別依變數，以及顯著值和自由度。如果是 CRT，則會顯示改善值。如果是 QUEST，則會為尺度與次序自變數顯示 F、顯著值與自由度；如果是名義自變數，則會顯示卡方值、顯著值與自由度。
- **節點定義。** 節點定義會顯示每個節點分割使用之自變數的值。

表格格式中的樹狀結構。 樹狀結構中每個節點的摘要資訊，包括父節點數、自變數統計量、節點的自變數值、尺度依變數的平均數與標準差，或是類別依變數的個數與百分比。

圖表 1-20
表格格式的樹(F)

	Bad		Good		總數				主要的自變數				
節點	N	百分比	N	百分比	N	百分比	預測的類別	父節點	變數	Sig.[a]	卡方	df	分隔值
0	1020	41.4%	1444	58.6%	2464	100.0%	Good						
1	454	82.1%	99	17.9%	553	22.4%	Bad	0	Income level	.000	662.457	2	<= Low
2	476	42.0%	658	58.0%	1134	46.0%	Good	0	Income level	.000	662.457	2	(Low, Medium]
3	90	11.6%	687	88.4%	777	31.5%	Good	0	Income level	.000	662.457	2	> Medium
4	422	56.7%	322	43.3%	744	30.2%	Bad	2	Number of credit cards	.000	193.113	1	5 or more
5	54	13.8%	336	86.2%	390	15.8%	Good	2	Number of credit cards	.000	193.113	1	Less than 5
6	80	17.6%	375	82.4%	455	18.5%	Good	3	Number of credit cards	.000	38.587	1	5 or more
7	10	3.1%	312	96.9%	322	13.1%	Good	3	Number of credit cards	.000	38.587	1	Less than 5
8	211	80.8%	50	19.2%	261	10.6%	Bad	4	Age	.000	95.299	1	<= 28.079
9	211	43.7%	272	56.3%	483	19.6%	Good	4	Age	.000	95.299	1	> 28.079

統計

圖表 1-21
「輸出」對話方塊，「統計量」索引標籤

可用的統計量表格根據依變數的測量水準、成長方法與其他設定而定。

模式

摘要。 摘要包括使用的方法、模式中所包括的變數，以及模式中所指定但未包括的變數。

圖表 1-22
模式摘要表

模式摘要

規格	成長方法	CHAID
	依變數	Credit rating
	自變數	Age, Income level, Number of credit cards, Education, Car loans
	有效性	無
	最大樹狀結構深度	3
	父節點中最小的觀察值	400
	子節點中最小的觀察值	200
結果	所包含的自變數	Income level, Number of credit cards, Age
	節點數量	10
	終端節點數量	6
	深度	3

風險。 風險估計與其標準誤。樹狀結構預測準確性的測量。

- 如果是類別依變數，風險估計就是在事前機率和錯誤分類成本調整之後，不正確分類之觀察值的比例。
- 如果是尺度依變數，風險估計是在節點變異數的範圍中。

分類表。 如果是類別（名義、次序）依變數，這個表格就會顯示為每個依變數類別正確分類與不正確分類之的觀察值數。不適用於尺度依變數。

圖表 1-23
風險和分類表

風險

估計	標準錯誤
.205	.008

成長方法: CHAID
依變數: Credit rating

分類

	預測的		
觀察的	Bad	Good	百分比修正
Bad	665	355	65.2%
Good	149	1295	89.7%
整體百分比	33.0%	67.0%	79.5%

成長方法: CHAID
依變數: Credit rating

成本、事前機率、分數與利潤值。 如果是類別依變數，這個表格會顯示分析中所使用的成本、事前機率、分數與利潤值。不適用於尺度依變數。

自變數

模式的重要性。 如果是 CRT 成長方法，則會根據其對模式的重要性來將每個自（預測值）變數分等。不適用於 QUEST 或 CHAID 方法。

根據分割來代理。 對於 CRT 與 QUEST 成長方法，如果模式包括代理，則會列出樹狀結構中每個分割的代理。不適用於 CHAID 方法。 如需詳細資訊，請參閱第 14 頁代理。

節點效能

摘要。 如果是尺度依變數，這個表格會包括節點數、觀察值數，以及依變數的平均值。如果是具有已定義之利潤的類別依變數，則表格會包括節點數、平均利潤以及 ROI（投資報酬率）值。不適用於沒有已定義之利潤的類別依變數。 如需詳細資訊，請參閱第 16 頁利潤。

圖表 1-24
節點與百分位數的獲利摘要表

節點增益摘要

節點	N	百分比	利潤	投資報酬率 (ROI)
7	322	13.1%	77.826	377.4%
5	390	15.8%	70.308	308.8%
6	455	18.5%	67.692	287.9%
9	483	19.6%	49.420	172.0%
8	261	10.6%	23.410	64.7%
1	553	22.4%	22.532	61.9%

百分比增益摘要

百分位數	節點	N	利潤	投資報酬率 (ROI)
10	7	246	77.826	377.4%
20	7; 5	493	75.218	352.0%
30	5; 6	739	73.488	336.2%
40	6	986	72.036	323.4%
50	6; 9	1232	70.205	307.9%
60	9	1478	66.745	280.6%
70	9; 8	1725	63.134	254.4%
80	8; 1	1971	58.149	221.6%
90	1	2218	54.183	197.9%
100	1	2464	51.023	180.4%

依目標分類。 如果是具有已定義之目標分類的類別依變數，則表格會包括百分比獲利、回應百分比以及根據節點或百分位數組別所區分的索引百分比（提升）。每個目標類別都會產生個別的表格。不適用於沒有已定義之目標類別的尺度依變數或類別依變數。 如需詳細資訊，請參閱第 5 頁選取類別。

圖表 1-25
節點或百分位數的目標類別獲利

目標類別: Bad

節點增益

節點	節點		得到		回應	索引
	N	百分比	N	百分比		
1	553	22.4%	454	44.5%	82.1%	198.3%
8	261	10.6%	211	20.7%	80.8%	195.3%
9	483	19.6%	211	20.7%	43.7%	105.5%
6	455	18.5%	80	7.8%	17.6%	42.5%
5	390	15.8%	54	5.3%	13.8%	33.4%
7	322	13.1%	10	1.0%	3.1%	7.5%

百分比增益

百分位數	節點	N	得到		回應	索引
			N	百分比		
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
50	9	1232	848	83.1%	68.8%	166.2%
60	9 ; 6	1478	908	89.0%	61.4%	148.4%
70	6	1725	951	93.3%	55.1%	133.2%
80	6 ; 5	1971	986	96.7%	50.0%	120.9%
90	5 ; 7	2218	1012	99.3%	45.6%	110.3%
100	7	2464	1020	100.0%	41.4%	100.0%

列。 節點效能表可以根據終端節點、百分位數或兩者來顯示結果。如果您選取同時使用兩者，每個目標類別就會產生兩個表格。百分位數表會根據排序順序，為每個百分位數顯示累積值。

百分位數增量。 如果是百分位數表，您可以選取百分位數增量：1、2、5、10、20 或 25。

顯示累積統計量。 如果是終端節點表，會在每個表格顯示更多的欄位，來顯示累積結果。

圖表

圖表 1-26
「輸出」對話方塊，「圖形」索引標籤

可用的圖表是根據依變數的測量水準、成長方法與其他設定而定。

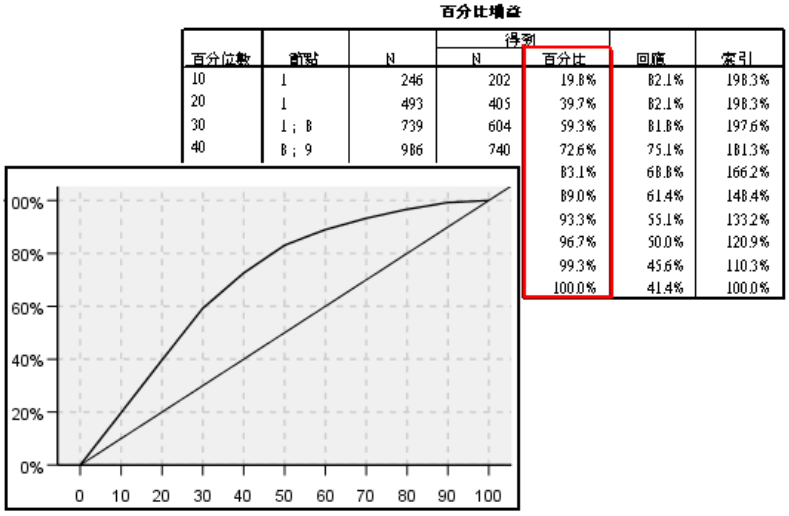
自變數對模式的重要性。 根據自變數（測量值）之模式重要性長條圖。只適合與 CRT 成長方法搭配使用。

節點效能

獲利。「獲利」是指每個節點之目標類別的總觀察值的百分比，計算方式為：(節點目標 n /總目標 n) x 100。獲利圖表就是累積百分位數獲利的線形圖，計算方式為：(累積百分位數目標 n / 總目標 n) x 100。會為每個目標類別產生個別的線性圖。只適用於有定義之目標類別的類別依變數。 如需詳細資訊，請參閱第 5 頁選取類別。

獲利圖表會繪製與您在百分位數表之獲利的「獲利百分比」行中所見相同的值，這個百分位數表也會報告累積值。

圖表 1-27
百分位數表與獲利圖表的獲利



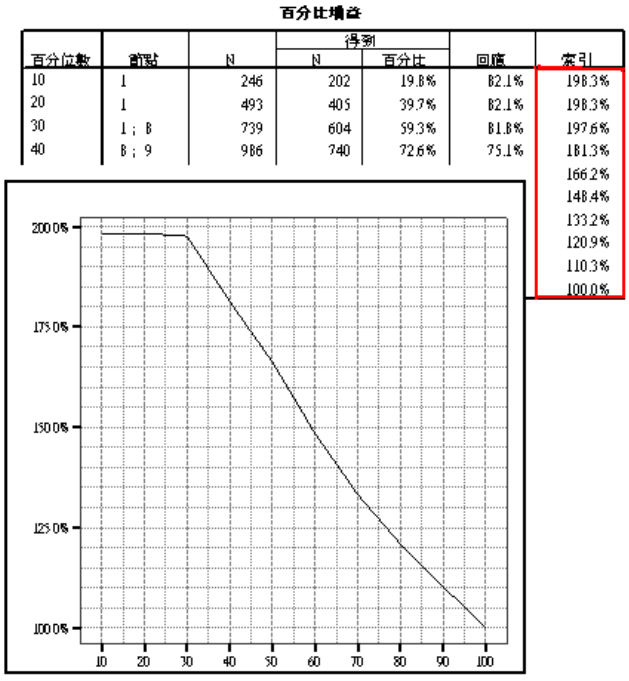
百分比增益

百分位數	節點	N	得到 N	得到 百分比	回應	索引
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
			831	83.1%	68.8%	166.2%
			890	89.0%	61.4%	148.4%
			933	93.3%	55.1%	133.2%
			967	96.7%	50.0%	120.9%
			993	99.3%	45.6%	110.3%
			1000	100.0%	41.4%	100.0%

指數。 指數為目標類別之節點回應值百分比與整個樣本之整體目標類別回應值百分比相較之下，所得出的比率。 索引圖表就是累積百分位數索引值的線性圖。僅適用於類別依變數。累積百分位數索引的計算方式為：（累積百分位數回應百分比 / 總回應百分比）x 100。會為每個目標類別產生個別的圖表，而且目標類別必須已經定義。

索引圖表會繪製與您在百分位數表之獲利的「索引」行中所見相同的值。

圖表 1-28
百分位數表與索引圖表的獲利



百分比增益

百分位數	節點	N	得到 N	得到 百分比	回應	索引
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
						166.2%
						148.4%
						133.2%
						120.9%
						110.3%
						100.0%

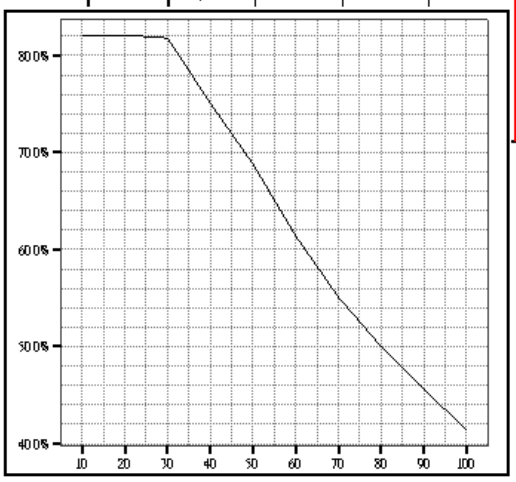
回應。 指定的目標類別中，節點內的觀察值百分比。 回應圖表就是累積百分位數回應的線性圖，計算方式為：（累積百分位數目標 n / 累積百分位數總數 n）x 100。僅適用於具有已定義之目標類別的類別依變數。

回應圖表會繪製與您在百分位數表之獲利的「回應」行中所見相同的值。

圖表 1-29
百分位數表與回應圖表的獲利

百分比增益

百分位數	節點	N	得到 N	得到 百分比	回應	索引
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
					68.8%	166.2%
					61.4%	148.4%
					55.1%	133.2%
					50.0%	120.9%
					45.6%	110.3%
					41.4%	100.0%



平均數。 依變數的累積百分位數平均值線性圖。僅適用於尺度依變數。

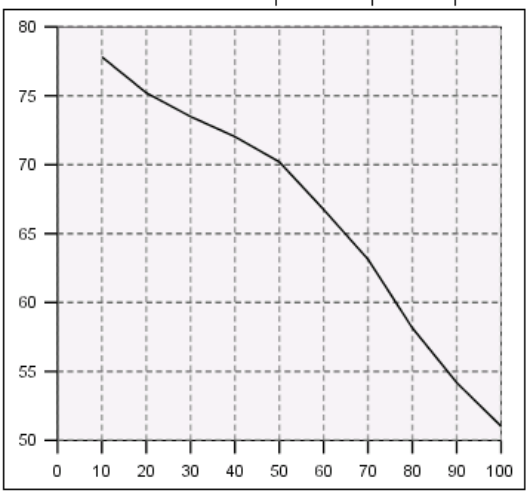
平均利潤。 累積平均利潤的線性圖。僅適用於具有已定義之利潤的類別依變數。
如需詳細資訊，請參閱第 16 頁利潤。

平均利潤圖表會繪製與您在百分位數表之獲利摘要的「利潤」行中所見相同的值。

圖表 1-30
百分位數表的獲利摘要與平均利潤圖表

百分比增益摘要

百分位數	截點	N	利潤	投資報酬率 (ROI)
10	7	246	77.826	377.4%
20	7; 5	493	75.218	352.0%
30	5; 6	739	73.488	336.2%
40	6	986	72.036	323.4%
			70.205	307.9%
			66.745	280.6%
			63.134	254.4%
			58.149	221.6%
			54.183	197.9%
			51.023	180.4%



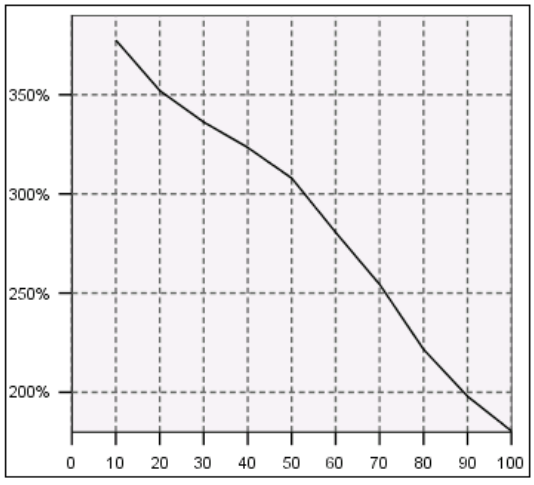
投資報酬率（ROI）。 累積的 ROI（投資報酬）的線性圖。ROI 是以利潤對支出的比率來計算。僅適用於具有已定義之利潤的類別依變數。

ROI 圖表會繪製與您在百分位數表之獲利摘要的「ROI」行中所見相同的值。

圖表 1-31
百分位數表的獲利摘要與 ROI 圖表

百分比增益摘要

百分位數	截點	N	利潤	投資報酬率 (ROI)
10	7	246	77.826	377.4%
20	7; 5	493	75.218	352.0%
30	5; 6	739	73.488	336.2%
40	6	986	72.036	323.4%
				307.9%
				280.6%
				254.4%
				221.6%
				197.9%
				180.4%



百分位數增量。 對於所有的百分位數圖表，這個設定會控制圖表上所顯示的百分位數增量：1、2、5、10、20 或 25。

選項與分數規則

圖表 1-32
「輸出」對話方塊，「規則」索引標籤

「規則」索引標籤會提供以指令語法、SQL 或範例（純英文）文字等形式來產生選項或分類/預測規則的功能。您可以在「瀏覽器」中顯示這些規則及/或將這些規則儲存至某個外部檔案。

語法。 控制在「瀏覽器」中顯示之輸出以及儲存為外部檔案兩者的選擇規則。

- **IBM SPSS Statistics.** 指令語法語言。規則是以定義用於選取觀察值子集之過濾條件的一組指令來表示，或以用於為觀察值評分的 COMPUTE 陳述式來表示。
- **SQL。** 標準的 SQL 規則是用來從資料庫中選取或擷取記錄，或是將值指定給這些記錄。產生的 SQL 規則不包含任何表格名稱或其他資料來源資訊。
- **簡單文字。** 純英文虛擬程式碼。規則是表示為一組邏輯 “if...then” 陳述式，這一組陳述式可以描述模式的分類或每一個節點的預測。這種形式的規則可以用來定義變數和數值標記或變數名稱和資料值。

類型。 若是 SPSS Statistics 和 SQL 規則，可控制所產生規則的類型：選擇或評分規則。

- **指定值給觀察值。** 此規則可用來指定模式的預測給符合節點成員資格條件的觀察值。另外會為符合節點成員資格條件的各節點產生不同的規則。
- **選取觀察值。** 此規則可用來選取符合節點成員資格條件的觀察值。有關 SPSS Statistics 或 SQL 規則，會產生單一規則，以選取符合選擇條件的所有觀察值。

SPSS Statistics 和 SQL 規則中包含代理。您可以在 CRT 和 QUEST 中，包含規則中模式的代理預測值。包含代理的規則可能會相當複雜。一般來說，如果只要推導有關樹狀結構的概念資訊，請排除代理。如果有些觀察值有不完整的自變數（預測值）資料，而您要模擬樹狀結構的規則，請包含代理。 如需詳細資訊，請參閱第 14 頁代理。

節點。 控制產生規則的範圍。個別的規則會針對範圍中所包括的每個節點而產生。

- **所有的終端節點。** 為每個終端節點產生規則。
- **最佳終端節點。** 根據索引值，為前 n 個終端節點產生規則。如果數量超過樹狀結構中終端節點的數量，就會為所有的終端節點產生規則(請參閱以下注意事項)。
- **最佳終端節點會往上移至指定的觀察值百分比。** 根據索引值，為前 n 百分比觀察值的終端節點產生規則(請參閱以下注意事項)。
- **其索引值符合或超過分割值的終端節點。** 為具有索引值大於或等於指定值的所有終端節點產生規則。大於 100 的索引值是表示該節點中目標類別內的觀察值百分比已經超過根節點中的百分比(請參閱以下注意事項)。
- **所有節點。** 為所有節點產生規則。

注意 1：以索引值為根據的節點選項功能僅適用於有已定義之目標類別的類別依變數。如果您已經指定多個目標類別，就會為每個目標類別產生一組個別的規則。

注意 2：如果是用於選擇觀察值的 SPSS Statistics 與 SQL 規則（不是用於指定值的規則），則「所有節點」與「所有終端節點」將會有效率地產生可以選擇分析中所使用之所有觀察值的規則。

將規則匯出至檔案。 將規則儲存在某個外部文字檔案中。

您也可以根據最終樹狀結構模式中已經選取的節點，以互動方式產生並儲存選項或分數規則。 如需詳細資訊，請參閱第 42 頁第 2 章中的觀察值選擇和評分規則。

注意：如果您將指令語法格式的規則套用到另一個資料檔案，則該資料檔案必須包含與最終模式中之自變數相同名稱、使用相同之單位測量，並且有使用者定義遺漏值（如果有的話）的變數。

章節
2

樹狀編輯器

使用「樹狀結構編輯程式」時，您可以：

- 隱藏和顯示選擇的樹狀結構分支。
- 控制節點內容、分割節點的統計量，以及其他資訊的顯示。
- 變更節點、背景、框線、圖表和字型顏色。
- 變更字型樣式和大小。
- 變更樹狀結構對齊方式。
- 依據選擇的節點，選擇要進一步分析的觀察值子集。
- 依據選擇的節點，建立和儲存選擇或評分觀察值的規則。

若要編輯樹狀結構模式：

▶ 在「瀏覽器」視窗中，連按兩下樹狀結構模式。

或

▶ 在「編輯」功能表或按一下滑鼠右鍵的快顯功能表上，請選擇：
編輯內容(O) > 在個別視窗中(W)

隱藏和顯示節點

若要隱藏（收合）父節點下分支中的所有子節點：

▶ 在父節點的右下角，按一下小方塊中的減號（-）。

在該分支父節點下所有的子節點將會隱藏。

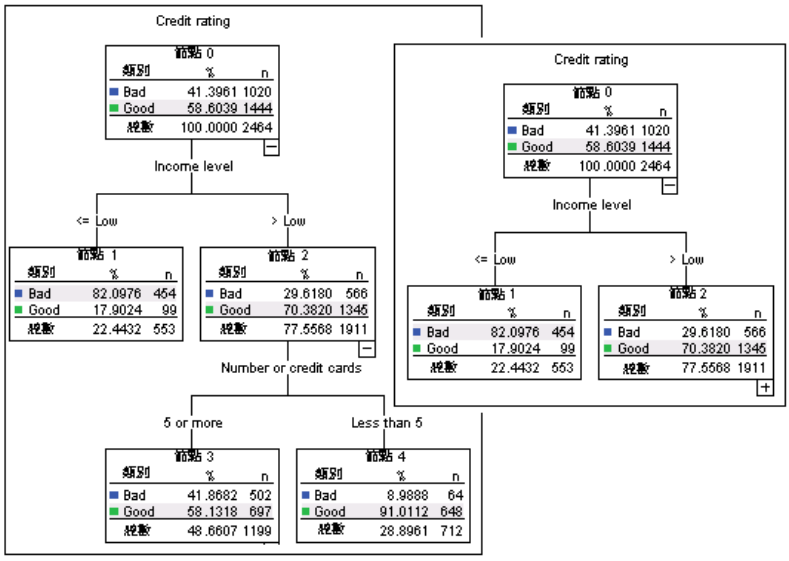
若要顯示（展開）父節點下分支中的所有子節點：

▶ 在父節點的右下角，按一下小方塊中的加號（+）。

注意：隱藏分支中的子節點與修正樹狀結構是不一樣的。如果您要的是已修正的樹狀結構，您必須在建立樹狀結構之前要求修正，而且已修正的分支不會包含在最後的樹狀結構之中。如需詳細資訊，請參閱第 13 頁第 1 章中的修正樹狀結構。

Copyright IBM Corporation 1989, 2011.

圖表 2-1
已展開和已收合的樹狀結構



選擇多個節點

您可以依據目前選擇的節點，選擇觀察值、產生評分和選擇規則，以及執行其他動作。若要選擇多個節點：

- 按一下您要選擇的節點。
- 按住 Ctrl 鍵不放，然後再按您要選擇的節點。

您可以選擇一個分支中多個相鄰的節點和（或）父節點，以及其他分支中的子節點。但是，您不能選擇同一個節點分支中的父節點和子節點/其下節點。

使用大型樹狀結構

有時候，樹狀結構模式包含有太多節點和分支，很難或甚至不可能檢視整個完整的樹狀結構。在使用大型樹狀結構時，有一些實用的功能：

- **樹狀結構圖。** 您可以使用樹狀結構圖（尺寸較小，是樹狀的簡化版）瀏覽樹狀結構和選擇節點。 如需詳細資訊，請參閱第 36 頁樹狀圖。
- **縮放比例。** 您可以變更縮放比例，縮小或放大樹狀結構顯示。 如需詳細資訊，請參閱第 37 頁縮放樹狀結構顯示。
- **節點和分支顯示。** 您可以利用只顯示節點中的表格或圖表，和（或）隱藏節點標記或自變數的顯示資訊，使樹狀結構看起來更精簡。 如需詳細資訊，請參閱第 39 頁控制樹狀結構中顯示的資訊。

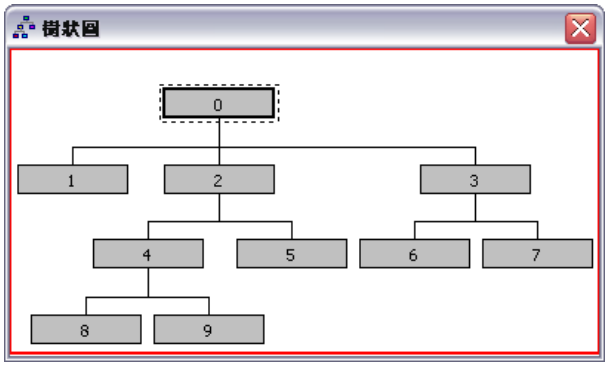
樹狀圖

樹狀結構圖提供精簡、簡化的樹狀結構檢視，讓您可以瀏覽樹狀結構和選擇節點。

若要使用樹狀結構圖視窗：

▶ 從「樹狀結構編輯程式」功能表選擇：
檢視 > 樹狀圖

圖表 2-2
樹狀結構圖視窗



- 目前選擇的節點會在「樹狀結構模式編輯程式」和樹狀結構圖視窗中反白顯示。
- 樹狀結構圖中的紅色方框表示目前正在「樹狀結構模式編輯程式」中檢視的區域。按一下滑鼠右鍵並拖曳方框可變更檢視區域中顯示的樹狀結構區段。
- 如果您在樹狀結構圖中選擇了目前不在「樹狀結構編輯程式」檢視區域中的節點，則檢視區域會移至包含該選取節點的區域以供檢視。
- 在樹狀結構圖和「樹狀結構編輯程式」中選擇多個節點的方式相同：按住 Ctrl 鍵不放，選擇多個節點。您不能選擇同一個節點分支中的父節點和子節點/其下節點。

縮放樹狀結構顯示

依照預設值，樹狀結構會自動縮放至符合「瀏覽器」視窗的大小，因此部分樹狀結構在剛開始時較不容易讀取。您可以選擇預設的縮放比例設定值，或是輸入您自訂的縮放比例值，範圍從 5% 到 200%。

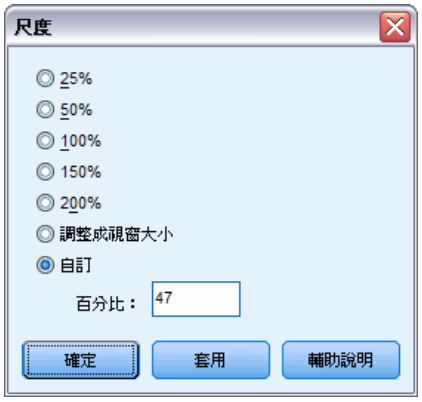
若要變更樹狀結構的縮放比例值：

▶ 在工具列的下拉式清單中，選擇縮放比例，或是輸入自訂的比例值。

或

▶ 從「樹狀結構編輯程式」功能表選擇：
檢視 > 尺度...

圖表 2-3
「縮放比例」對話方塊



您可以在建立樹狀結構模式前指定縮放比例值。 如需詳細資訊，請參閱第 22 頁第 1 章中的輸出。

節點摘要視窗

節點摘要視窗提供已選擇節點的放大檢視。您也可以依據選擇節點，使用摘要視窗來檢視、套用，或是儲存選項或評分規則。

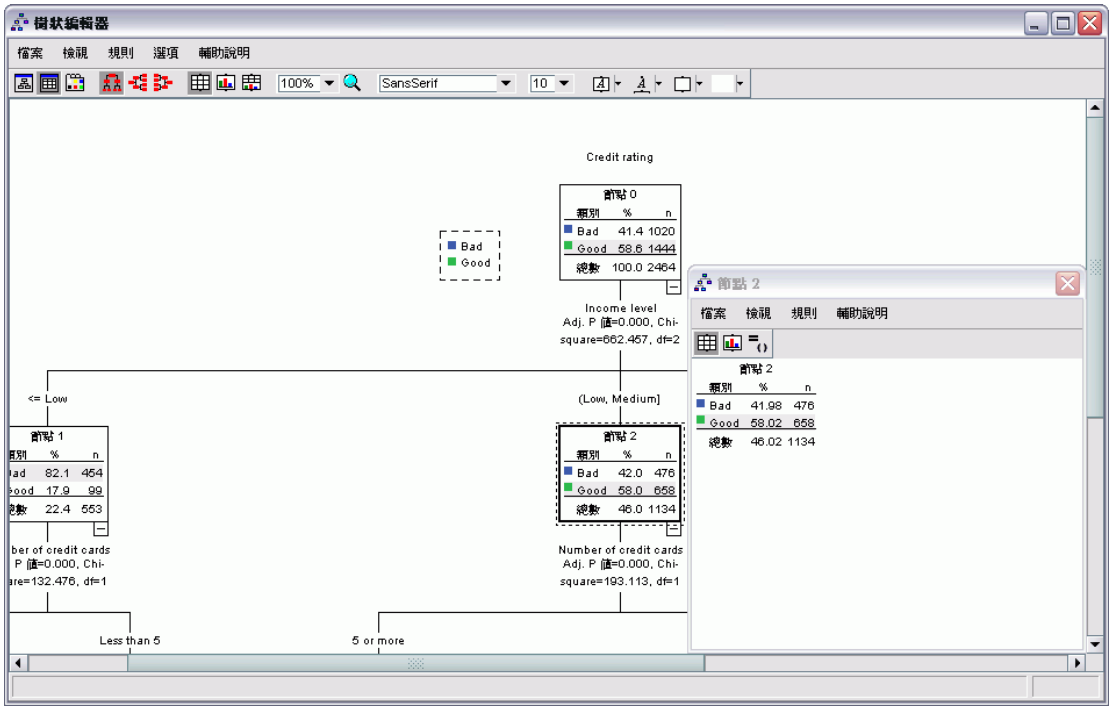
- 在節點摘要視窗中，使用「檢視」功能表切換檢視摘要表格、圖表或規則。
- 在節點摘要視窗中，使用「規則」功能表選擇您要查看的規則類型。 如需詳細資訊，請參閱第 42 頁觀察值選擇和評分規則。
- 所有節點摘要視窗中的檢視會反映所有已選擇節點的組合摘要。

若要使用節點摘要視窗：

▶ 在「樹狀結構編輯程式」中選擇節點。若要選擇多個節點，可以按住 Ctrl 鍵來選取。

▶ 從功能表選擇：
檢視 > 摘要

圖表 2-4
摘要視窗

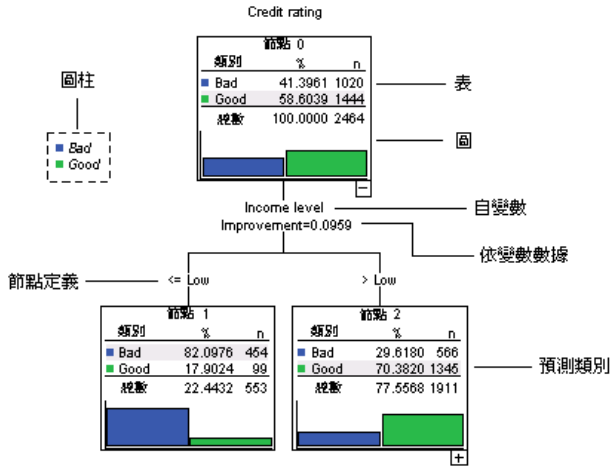


控制樹狀結構中顯示的資訊

「樹狀結構編輯程式」中的「選項」功能表可讓您控制顯示節點內容、自變數（預測變數）名稱和統計量、節點定義和其他設定值。其中許多設定值也可以從工具列進行控制。

設定	選項功能表的選項
反白顯示預測類別（類別依變數）	預測重要性
節點中的表格和（或）圖表	節點內容
顯著性檢定值和 p 值	自變數統計量
自變數（預測變數）名稱	自變數
節點的自變數（預測變數）值	節點定義
對齊（由上至下、由左至右、由右至左）	方向
圖表圖註	圖註

圖表 2-5
樹狀結構元素



變更樹狀結構的顏色和字型

您可以在樹狀結構中變更如下的顏色：

- 節點框線、背景和文字顏色
- 分支顏色和分支文字顏色
- 樹狀結構背景顏色
- 預測類別反白顯示的顏色（類別依變數）
- 節點圖表顏色

您可以變更樹狀結構中所有的字型、樣式和大小。

注意：您無法變更個別節點或分支的顏色或字型屬性。顏色變更會套用至所有相同類型的元素，以及字型變更（不同於顏色）會套用至所有圖表的元素。

若要變更顏色和字型屬性：

▶ 使用工具列變更整個樹狀結構的字型屬性，或是不同樹狀結構元素的顏色(當您將滑鼠游標移至工具列的控制項上方，「工具提示」會顯示說明資訊)。

或

▶ 在「樹狀結構編輯程式」的任意處連按兩下開啟「性質」視窗，或是在功能表中選擇：**檢視** > **內容**

▶ 有關框線、分支、節點背景、預測類別，和樹狀結構背景，按一下「**顏色**」索引標籤。

▶ 有關字型顏色和屬性，按一下「**文字**」索引標籤。

▶ 有關節點圖表顏色，按一下「**節點圖表**」索引標籤。

圖表 2-6
「性質」視窗，「顏色」索引標籤

圖表 2-7
「性質」視窗，「文字」索引標籤

圖表 2-8
「性質」視窗，「節點圖表」索引標籤

觀察值選擇和評分規則

您可以利用「樹狀結構編輯程式」，執行下列動作：

- 依據選擇的節點，選擇觀察值子集。 如需詳細資訊，請參閱第 42 頁過濾觀察值。
- 產生IBM® SPSS® Statistics指令語法或 SQL 格式的觀察值選擇或評分規則。 如需詳細資訊，請參閱第 43 頁儲存選擇和評分規則。

當您執行「決策樹狀結構」程序來建立樹狀結構模式時，您也可以依據多個準則自動儲存規則。 如需詳細資訊，請參閱第 33 頁第 1 章中的選項與分數規則。

過濾觀察值

如果您想要進一步瞭解特定節點或節點群組中的觀察值，您可以依據選擇的節點來選取要進一步分析的觀察值子集。

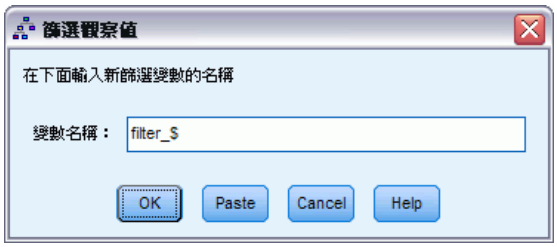
▶ 在「樹狀結構編輯程式」中選擇節點。若要選擇多個節點，可以按住 Ctrl 鍵來選取。

▶ 從功能表選擇：
規則 > **篩選觀察值...**

▶ 輸入過濾變數名稱。選擇節點中的觀察值將收到變數值 1。所有其他觀察值將會收到數值 0，並將在接下來的分析中被執行，直到您變更過濾狀態為止。

▶ 按一下「**確定**」。

圖表 2-9
「過濾觀察值」對話方塊



儲存選擇和評分規則

您可以將觀察值選擇或評分規則儲存在外部檔案，然後套用那些規則至不同的資料來源。這些規則是依據「樹狀結構編輯程式」中選擇的節點。

語法。 控制在「瀏覽器」中顯示之輸出以及儲存為外部檔案兩者的選擇規則。

- **IBM SPSS Statistics.** 指令語法語言。規則是以定義用於選擇觀察值子集之過濾條件的一組指令來表示，或以用於為觀察值計分的 COMPUTE 陳述式來表示。
- **SQL。** 標準 SQL 規則是用來從資料庫中選擇/擷取記錄，或指定值給這些記錄。產生的 SQL 規則不包含任何表格名稱或其他資料來源資訊。

類型。 您可以建立選擇或評分規則。

- **選擇觀察值。** 此規則可用來選擇符合節點成員資格條件的觀察值。有關 SPSS Statistics 或 SQL 規則，會產生單一規則，以選擇符合選擇條件的所有觀察值。
- **指定值給觀察值。** 此規則可用來指定模式的預測給符合節點成員資格條件的觀察值。另外會為符合節點成員資格條件的各節點產生不同的規則。

包括代理。 您可以在 CRT 和 QUEST 中，包含規則中模式的代理預測值。包含代理的規則可能會相當複雜。一般來說，如果只要推導有關樹狀結構的概念資訊，請排除代理。如果有些觀察值有不完整的自變數（預測值）資料，而您要模擬樹狀結構的規則，請包含代理。 如需詳細資訊，請參閱第 14 頁第 1 章中的代理。

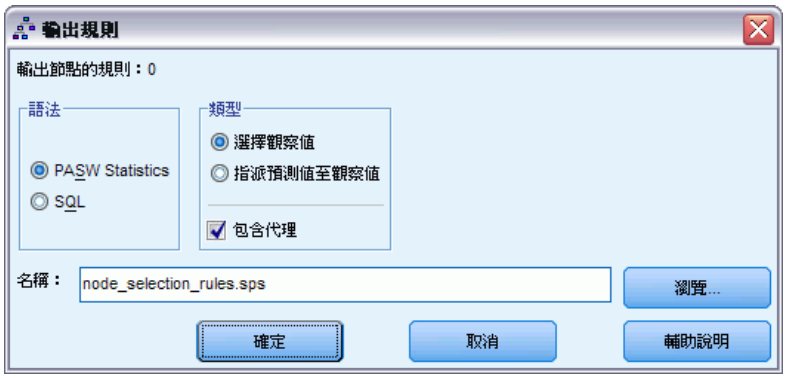
若要儲存觀察值選擇或評分規則：

▶ 在「樹狀結構編輯程式」中選擇節點。若要選擇多個節點，可以按住 Ctrl 鍵來選取。

▶ 從功能表選擇：
規則 > 輸出...

▶ 選擇您需要的規則類型，然後輸入檔名。

圖表 2-10
「匯出規則」對話方塊



注意：如果您將指令語法格式的規則套用到另一個資料檔案，則該資料檔案必須包含與最終模式中之自變數相同名稱、使用相同之單位測量，並且有使用者定義遺漏值（如果有的話）的變數。

部 II:
範例

章節
3

資料假設和需求

「決策樹狀結構」程序假設：

- 所有的分析變數已指派適當測量水準。
- 對於類別（**名義**、**次序**）依變數，分析中應包括的所有類別均已定義數值標記。

我們將使用檔案 tree_textdata.sav 來說明這些需求的重要性。此資料檔案反映在定義任何屬性（例如測量水準或數值標記）之前，讀取或輸入資料之預測狀態。 如需詳細資訊，請參閱附錄 A 中的範例檔案中的IBM SPSS Decision Trees 20。

樹狀結構模式的測量水準作用

此資料檔中的兩個變數皆為數值，且已為這兩個變數指派**尺度**測量水準。但是（如我們稍後所見）這兩個變數都是真正的類別變數，使用數值代碼來代表類別值。

▶ 若要執行「決策樹狀結構」分析，請從功能表選擇：
分析(A) > 分類 > 樹...

在來源變數清單中，這兩個變數旁的圖示代表著它們被視為尺度變數。

圖表 3-1
具有兩個類別變數的「決策樹狀結構」主對話方塊

Copyright IBM Corporation 1989, 2011.

- 選取「dependent」作為依變數。
- 選取「independent」作為自變數。
- 按一下「**確定**」執行程序。
- 再次開啟「決策樹狀結構」對話方塊，按一下[**重設**]。
- 在來源清單中的「dependent」上按一下滑鼠右鍵，並選取內容功能表中的「**名義**」。
- 對來源清單中的變數 independent 執行相同的程序。

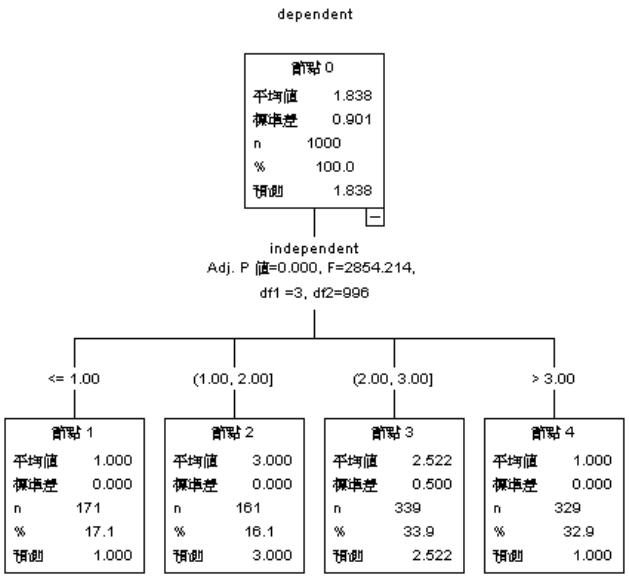
現在每個變數旁的圖示表示它們被視為名義變數。

圖表 3-2
來源清單中的名義圖示

- 選取「dependent」作為依變數，「independent」作為自變數，並按一下「**確定**」再次執行程序。

現在讓我們比較兩個樹狀結構。首先，我們將檢視將這兩個數值變數都視為尺度變數的樹狀結構。

圖表 3-3
兩個變數均視為尺度量數的樹狀結構

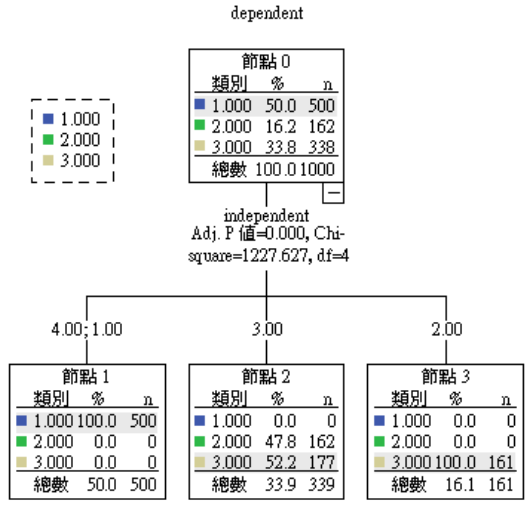


- 樹狀結構的每個節點均顯示「預測值」，這是該節點中依變數的平數值。對於真正為類別的變數，平均值是沒有意義的統計量。
- 該樹狀結構有四個子節點，每個節點分別代表每個依變數的數值。

樹狀結構模式通常會將類似的節點合併，但對於尺度變數，只會合併連續數值。在此範例中，沒有連續數值會視為相似到足以與任何節點合併在一起。

將兩個變數視為名義量數的樹狀結構在數個方面上有些不同。

圖表 3-4
兩個變數均視為名義量數的樹狀結構



- 每個節點會包含一個次數分配表（而非預測值），其中顯示依變數每個類別的觀察值數目（個數和百分比）。
- 「預測的」類別—每個節點中個數最多的類別—會加以反白。例如，節點 2 的預測類別為類別 3。
- 在此只有三個而非四個子節點，其中兩個自變數的數值合併到單一節點中。

合併到單一節點的兩個自變數為 1 和 4。因為根據定義，沒有繼承名義數值的順序，所以允許合併不連續的變數。

永久指派測量水準

當您在「決策樹狀結構」對話方塊中變更變數的測量水準時，變更只是暫時的，不會儲存至資料檔案。此外，您可能不會永遠知道什麼才是所有變數的正確測量水準。

「定義變數性質」可幫助您判斷出每個變數的正確測量水準，並永久變更指派的測量水準。若要使用「定義變數性質」：

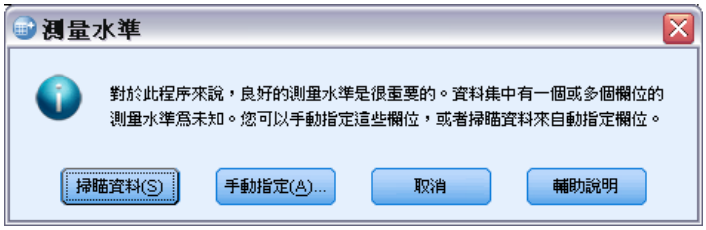
▶ 從功能表選擇：

資料 > 定義變數性質(V)...

具有未知測量水準的變數

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 3-5
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。

樹狀結構模式的數值標記作用

「決策樹狀結構」對話方塊介面假設類別（名義、次序）依變數的所有非遺漏值均已定義數值標記，或者沒有一個非遺漏值已定義。 有些功能至少需要類別依變數的兩個遺漏值具有數值標記，否則無法使用。如果至少兩個非遺漏值已定義數值標記，當有任何觀察值具有其他無數值標記的數值時，該觀察值會從分析中排除。

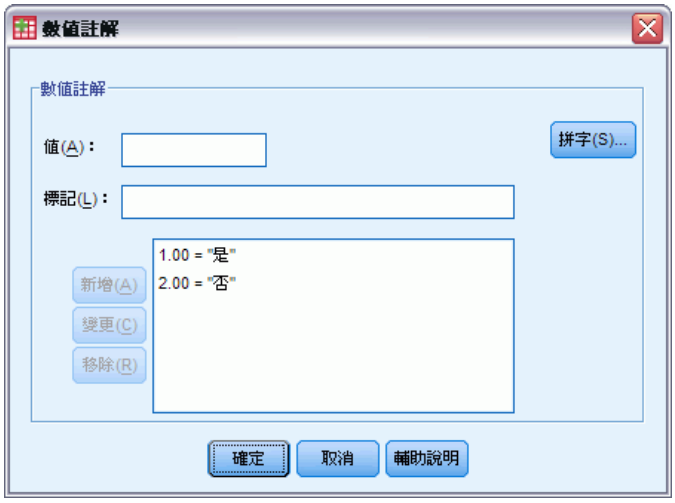
此範例中的原始資料檔沒有包含已定義數值標記，當依變數視為名義變數時，樹狀結構模式會在分析中使用所有未遺漏的數值。在此例中，這些數值為 1、2 和 3。

但是當我們為依變數的部分數值（而非有所數值）定義數值標記時，會發生什麼事？

▶ 在「資料編輯程式」視窗中，按一下「**變數檢視**」索引標籤。

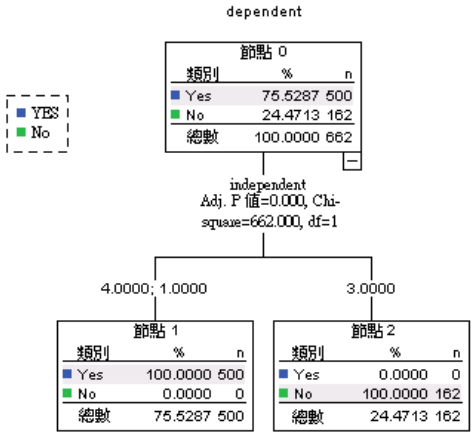
▶ 按一下變數「dependent」的「數值」儲存格。

圖表 3-6
定義依變數的數值標記



- 首先，在「數值」中輸入 1，在「數值標記」中輸入**是**，再按一下「**新增**」。
- 接下來，在「數值」中輸入 2，在「數值標記」中輸入**是**，再按一下「**新增**」。
- 然後按一下「**確定**」。
- 再次開啟「決策樹狀結構」對話方塊。對話方塊應仍選取 dependent 作為依變數，並具有名義測量水準。
- 按一下「**確定**」再次執行程序。

圖表 3-7
含部分數值標記的名義依變數樹狀結構



現在在樹狀結構模式中，只有兩個已定義數值標記的依變數。所有依變數值為 3 的觀察值已排除，如果您對資料不熟悉的話，可能不會很快察覺。

將數值標記指派給所有數值

若要避免在分析中不心遺漏了有效類別數值，請使用「定義變數性質」，將數值標記指派至在資料中找到的所有依變數值。

當變數 name 的資料字典資訊顯示在「定義變數性質」對話方塊中，您可以看到雖然該變數值為 3 的觀察值超過 300 個，該數值並未定義任何的數值標記。

圖表 3-8
「定義變數性質」對話方塊中，具有部分數值標記的變數

章節
4

使用決策樹狀結構來評估信用風險

銀行對於向銀行貸款的客戶，會建立一個客戶歷史資訊的資料庫，其中包括他們償還或拖欠貸款的紀錄。使用樹狀結構模式，您可以分析兩組客戶的特性，並建立出一個模式來預測貸款申請人會拖欠貸款的可能性。

信用資料儲存在 tree_credit.sav 中。 如需詳細資訊，請參閱附錄 A 中的範例檔案中的IBM SPSS Decision Trees 20。

建立模式

「決策樹狀結構程序」提供數種不同的方法，可用來建立樹狀結構模式。對於此例，我們會使用預設的方法：

CHAID。 卡方自動交互作用偵測。CHAID 會在每個步驟中，選擇與依變數具有最強交互作用的自（預測）變數。若與相關的依變數沒有明顯不同，則會合併每個預測變數的類別。

建構 CHAID 樹狀結構模式

▶ 若要執行「決策樹狀結構」分析，請從功能表選擇：
分析(A) > 分類 > 樹...

Copyright IBM Corporation 1989, 2011.

圖表 4-1
「決策樹狀結構」對話方塊

- 選取「信用評比」作為依變數。
- 選取所有其他的變數作為自變數。(此程序會自動排除任何對最終模式沒有顯著貢獻的變數)。

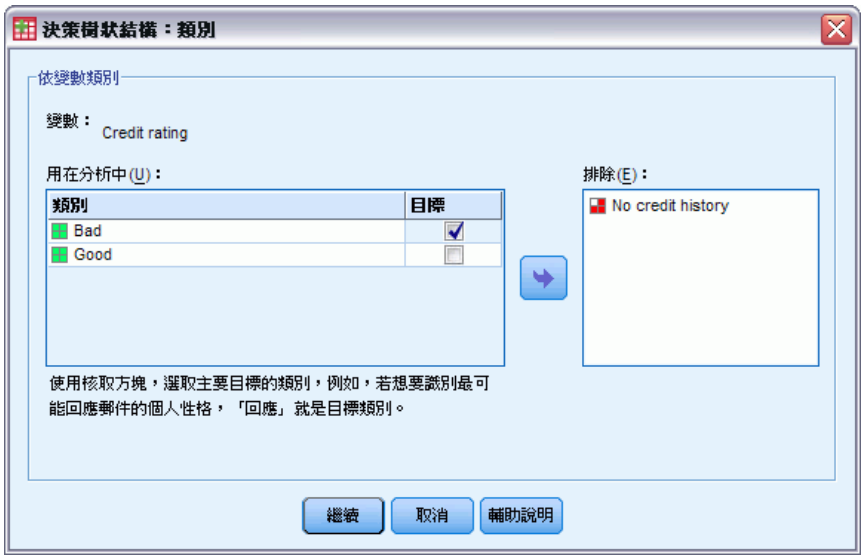
  此時，您可以執行程序，並產生基本樹狀結構模式，但我們要繼續選取一些額外的輸入值，並對用來產生模式的條件進行微幅調整。

選取目標類別

- 按一下所選依變數正下方的「類別」按鈕。

這會開啟「類別」對話方塊，在此您可以指定您所需的依變數目標類別。目標類別不影響樹狀結構模式本身，但有些輸出和選項只有在您選取目標類別後才能使用。

圖表 4-2
「類別」對話方塊



- 選取（勾選）「不良」類別的「目標」核取方塊。信用評比不良（會拖欠貸款）的客戶會視為所需的目標類別。
- 按一下「繼續」。

指定樹狀結構成長條件

就此例而言，我們要保持樹狀結構十分的簡單，所以我們要增加父節點和子節點觀察值的最小值，來限制樹狀結構的成長。

- 在主「決策樹狀結構」對話方塊中，按一下「條件」。

圖表 4-3
「條件」對話方塊，「成長限制」索引標籤

- 在「最小觀察值數」組別中，輸入 400 作為「父節點」，輸入 200 作為「子節點」。
- 按一下「繼續」。

選取額外的輸出。

- 在主「決策樹狀結構」對話方塊中，按一下「輸出」。

這會開啟有索引標籤的對話方塊，在此您可以選取各種類型的額外輸出。

圖表 4-4
「輸出」對話方塊，「樹狀結構」索引標籤

- 在「樹狀結構」索引標籤中，選取（勾選）「樹狀結構 - 表格格式」。
- 然後按一下「圖形」索引標籤。

圖表 4-5
「輸出」對話方塊，「圖形」索引標籤

- 選取（勾選）「**增益項**」和「**指數**」。

  注意：這些圖表需要依變數的目標類別。在此例中，在您指定一或多個目標類別後，才能存取「圖形」索引標籤。

- 按一下「**繼續**」。

儲存預測值

您可以儲存包含模式預測相關資訊的變數。例如，您可以儲存為每個觀察值預測的信用評比，再將這些預測與實際的信用評比比較。

- 在主「決策樹狀結構」對話方塊中，按一下「**儲存**」。

圖表 4-6
「儲存」對話方塊

- 選取（勾選）「終端節點數」、「預測值」和「預測機率」。
- 按一下「繼續」。
- 在主「分類樹狀結構」對話方塊中，按一下「確定」以執行程序。

評估模式

對於此例，模式結果包括：

- 提供模式相關資訊的表格。
- 樹狀結構圖。
- 提供模式效能指標的圖表。
- 新增至作用中資料集的模式預測變數。

模式摘要表

圖表 4-7
模式摘要(M)

模式摘要

規格	成長方法	CHAID
	依變數	Credit rating
	自變數	Age, Income level, Number of credit cards, Education, Car loans
	有效性	無
	最大樹狀結構深度	3
	父節點中最小的觀察值	400
	子節點中最小的觀察值	200
結果	所包含的自變數	Income level, Number of credit cards, Age
	節點數量	10
	終端節點數量	6
	深度	3

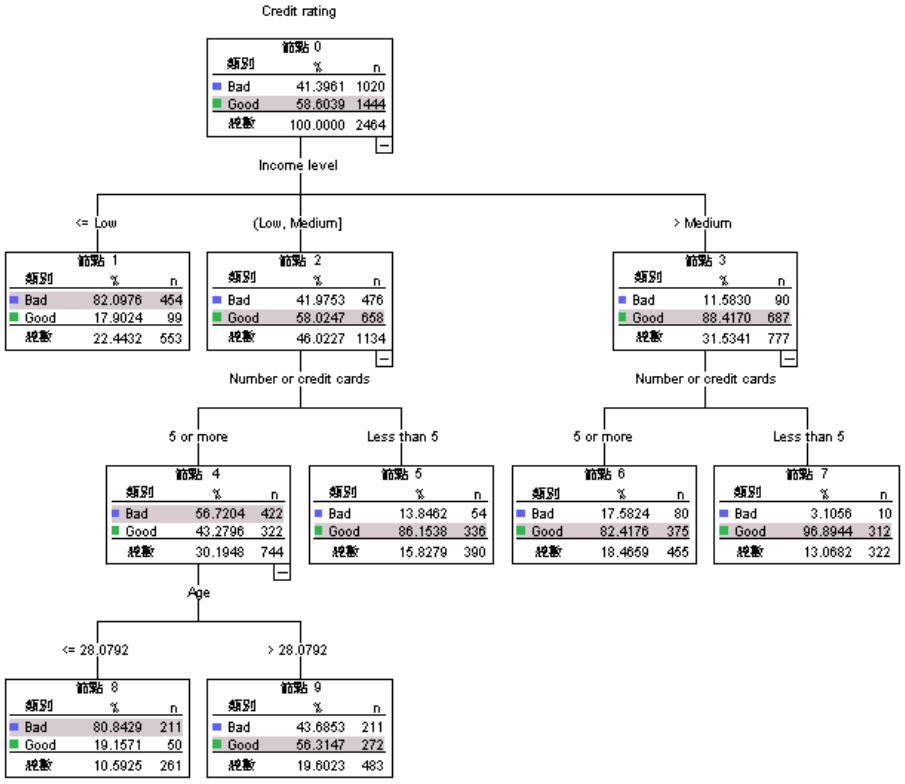
模式摘要表提供廣範圍的資訊，包括用來建構模式的規格與結果模式等相關資訊。

- 在「規格」區中，提供用來產生樹狀的規格模式的設定，其中包括在分析中使用的變數。
- 「結果」區顯示的資訊包括終端節點的總數、樹狀的規格深度（根節點下的層級數），和最終模式中的自變數。

在此已指定了五個自變數，但在最終模式中只會包含三個變數。教育變數和目前車貸變數的個數對模式沒有顯著的貢獻，所以它們會自動從最終模式中剔除。

樹狀結構圖

圖表 4-8
信用評比模式的樹狀結構圖



樹狀結構圖是樹狀結構模式的圖形表示。此樹狀結構圖顯示：

- 使用 CHAID 方式，收入水準是信用評比的最佳預測值。
- 對於低收入類別，收入水準是信用評比的唯一顯著預測值。在此類別的銀行客戶中，有 82% 拖欠貸款。由於其下沒有子節點，所以此類別視為**終端**節點。
- 對於中高收入類別，次佳的預測值是信用卡數目。
- 對於有五張以上（含）信用卡的中等收入客戶，模式包括一個以上的預測值：年齡。28 歲或以下的客戶超過 80% 有不良的信用評比，而超過 28 歲的客戶，有略不到一半的人有不良信用評比。

您可以使用「樹狀結構編輯程式」隱藏和顯示所選的分支、變更色彩和字型、並根據所選的節點選擇觀察值子集。 如需詳細資訊，請參閱第 67 頁選取節點中的觀察值。

樹狀結構表

圖表 4-9
信用評比的樹狀結構表

	Bad		Good		總數			
節點	N	百分比	N	百分比	N	百分比	預測的類別	父節點
0	1020	41.4%	1444	58.6%	2464	100.0%	Good	
1	454	82.1%	99	17.9%	553	22.4%	Bad	0
2	476	42.0%	658	58.0%	1134	46.0%	Good	0
3	90	11.6%	687	88.4%	777	31.5%	Good	0
4	422	56.7%	322	43.3%	744	30.2%	Bad	2
5	54	13.8%	336	86.2%	390	15.8%	Good	2
6	80	17.6%	375	82.4%	455	18.5%	Good	3
7	10	3.1%	312	96.9%	322	13.1%	Good	3
8	211	80.8%	50	19.2%	261	10.6%	Bad	4
9	211	43.7%	272	56.3%	483	19.6%	Good	4

樹狀結構表就如同它的名稱一樣，以表格形式提供大部分的必要樹狀結構資訊。對於每個節點，表格顯示：

- 在依變數每個類別中的觀察值數目和百分比。
- 依變數的預測類別。在此例中，由於只有兩個可能的信用評比，所以預測類別是信用評比類別，它在該節點中有超過 50% 的觀察值。
- 樹狀結構中每個節點的父節點。請注意，節點 1—低收入水準節點—不是任何節點的父節點。因為它是終端節點，所以它沒有子節點。

圖表 4-10
信用評比的樹狀結構表（續）

	主要的自變數			
變數	Sig.	卡方	df	分隔值
Income level	.000	662.457	2	<= Low
Income level	.000	662.457	2	(Low, Medium]
Income level	.000	662.457	2	> Medium
Number of credit cards	.000	193.113	1	5 or more
Number of credit cards	.000	193.113	1	Less than 5
Number of credit cards	.000	38.587	1	5 or more
Number of credit cards	.000	38.587	1	Less than 5
Age	.000	95.299	1	<= 28.079
Age	.000	95.299	1	> 28.079

- 用來分割節點的自變數。
- 分割用的卡方值（因為樹狀結構是以 CHAID 方法產生的）、自由度 (df)、和顯著性水準 (Sig.)。為達最實用的目的，您或許只會對顯著性水準有興趣；此模式中所有分割的顯著性水準均小於 0.0001。
- 該節點的自變數值。

注意：對於次序和尺度自變數，您會在樹狀結構和樹狀結構表中看到以一般格式 (value1, value2] 所表示的範圍，這基本上表示「大於 value1，小於或等於 value2」。 在此例中，收入水準只有三個可能值：—低、中 和 高—，(低, 中] 即代表中。類似的型式，>中表示高。

節點增益

圖表 4-11
節點增益

節點增益

節點	節點		得到		回應	索引
	N	百分比	N	百分比		
1	553	22.4%	454	44.5%	82.1%	198.3%
8	261	10.6%	211	20.7%	80.8%	195.3%
9	483	19.6%	211	20.7%	43.7%	105.5%
6	455	18.5%	80	7.8%	17.6%	42.5%
5	390	15.8%	54	5.3%	13.8%	33.4%
7	322	13.1%	10	1.0%	3.1%	7.5%

成長方法 CHAID
依變數: Creditrating

節點增益表提供有關模式中終端節點的摘要資訊。

- 在此表格中只會列出終端節點—樹狀結構停止成長—的節點。通常，您只會對終端節點有興趣，因為它們代表模式的最佳分類預測。
- 由於增益值提供有關目標類別的資訊，所以只有在您指定一或多個目標類別時，才能使用此表格。在此例中只有一個目標類別，所以只有一個節點增益表。
- 節點 N 是每個終端節點的觀察值數，節點百分比是每個節點中觀察值總數的百分比。
- 增益 N是目標類別中每個終端節點的觀察值，增益百分比是目標類別中觀察值的百分比，這與目標類別觀察值的總數有關 — 在此例中，為不良信用評比的觀察值數目和百分比。
- 對於類別依變數，回應是指定目標類別中節點內的觀察值百分比。在此例中，這些百分比是樹狀結構中不良類別所顯示的相同百分比。
- 對於類別依變數，指數是目標類別的回應百分比，對整個樣本的回應百分比之比率。

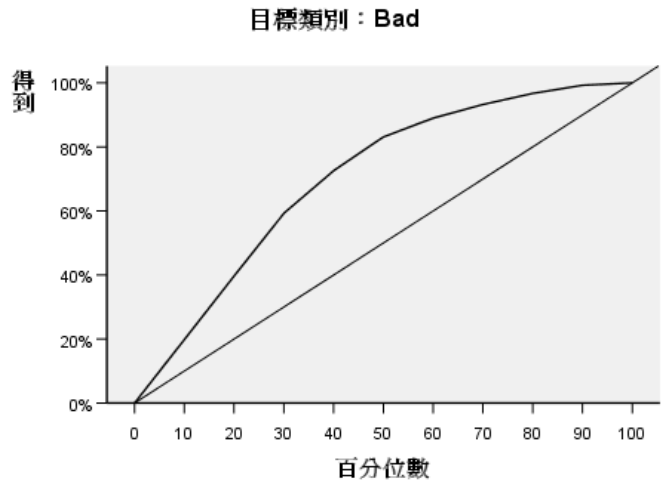
指標數值

指數值基本上是一種指標，用來表示該節點的觀察目標類別百分比，與期望目標類別百分比之間差異的程度。 根節點中的目標類別百分比代表採用任何自變數效果之前的期望百分比。

指數值大於 100%，表示目標類別中的觀察值大於目標類別的總百分比。 相反地，指數值小於 100%，表示目標類別中的觀察值小於目標類別的總百分比

增益圖表

圖表 4-12
不良信用評比目標類別的增益圖表

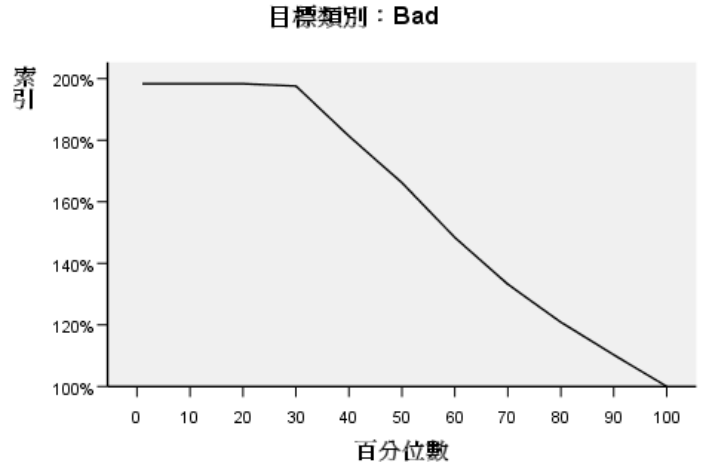


此增益圖表表示該模式是非常好的模式。

累計增益圖表起點一律為 0%，終點一律為 100%，如同您從一端移至另一端。對於一個好的模型，增益項表會陡然增加到 100%，然後維持水平。沒有提供資訊的模式，圖表會沿著對角參考線增加。

指數圖表

圖表 4-13
不良信用評比目標類別的指數圖表



此指數圖表指示該模式是好的模式。累計指數圖表的起點傾向大於 100%，然後逐漸減少一直到達 100% 為止。

對於一個好的模式，指數值的起點應大於 100%，然後維持在穩定的高原期，隨著您的移動，尾部會急遽地朝向 100% 停止。對於沒有提供資訊的模式，整個圖表的線條會在接近 100% 處盤旋。

風險估計和分類

圖表 4-14
風險和分類表

風險

估計	標準錯誤
.205	.008

成長方法: CHAID
依變數: Creditrating

分類

	預測的		
觀察的	Bad	Good	百分比修正
Bad	665	355	65.2%
Good	149	1295	89.7%
整體百分比	33.0%	67.0%	79.5%

成長方法: CHAID
依變數: Creditrating

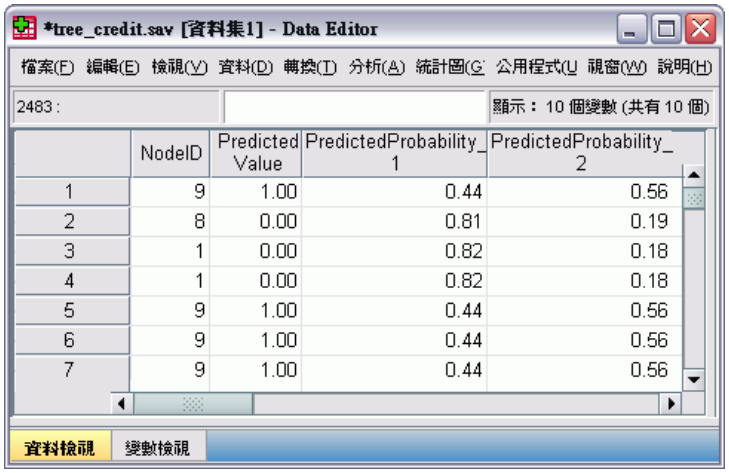
風險和分類表可快速評估模式的適用程度。

- 若風險估計為 0.205，表示該模式（優良或不良風險評比）所預測的類別，對 20.5% 的觀察值是錯誤的。所以將客戶錯誤分類的「風險」接近 21%。
- 分類表的結果與風險估計一致。表格顯示模型將大約 79.5% 的客戶分類正確。

但是，分類表確實顯示出此模式中一個潛在的問題：對於那些不良信用評比的客戶，它只能預測這類客戶中 65% 有不良評比，這表示有 35% 的不良信用評比客戶會錯誤地分類到「優良」客戶。

預測值

圖表 4-15
新的預測值和機率變數



	NodeID	PredictedValue	PredictedProbability_1	PredictedProbability_2
1	9	1.00	0.44	0.56
2	8	0.00	0.81	0.19
3	1	0.00	0.82	0.18
4	1	0.00	0.82	0.18
5	9	1.00	0.44	0.56
6	9	1.00	0.44	0.56
7	9	1.00	0.44	0.56

在作用中資料集中，已建立了四個新變數：

NodeID。 每個觀察值的終端節點數。

PredictedValue。 每個觀察值的依變數預測值。由於依變數的代碼 0 = 不良，1 = 優良，所以預測值為 0 表示該觀察值預測為不良信用評比。

PredictedProbability。 觀察值歸入到依變數每個類別的機率。因為依變數只有兩個可能值，所以建立兩個變數：

- **PredictedProbability_1。** 觀察值歸入不良信用評比的機率。
- **PredictedProbability_2。** 觀察值歸入優良信用評比的機率。

預測機率只是包含每個觀察值的終端節點，其每個依變數類別中的觀察值比例。例如，在節點 1 中，82% 的觀察值是不良類別，18% 是優良類別，預測的機率分別是 0.82 和 0.18。

對於類別依變數，每個觀察值的預測值是終端節點中，具有最高觀察值比例的類別。例如，對於第一個觀察值，預測值是 1（優良信用評比），因為在其終端節點中，約有 56% 的觀察值有優良信用評比。 相反的，對於第二個觀察值，預測值是 0（不良信用評比），因為在其終端節點中，約有 81% 的觀察值有不良信用評比。

但是如果您定義了成本，那麼預測類別和預測機率之間的關係可能沒有如此的直接。如需詳細資訊，請參閱第 70 頁指定成本至結果。

精確化模式

整體而言，模式的正確分類比率恰好低於 80%。這反映在多數的終端節點中，在此對於 80% 或以上的觀察值，預測類別 — 節點中反白的類別 — 與實際類別相同。

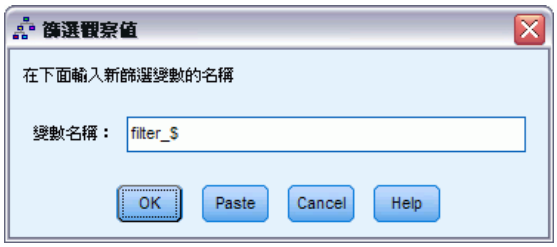
但是，有一個終端節點其觀察值非常平均地在優良和不良信用評比間分割。在節點 9 中，預測的信用評比是「優良」，但節點中只有 56% 的觀察值有優良信用評比。 這表示該節點幾乎一半的觀察值 (44%) 的預測類別是錯誤的。如果主要的考量是要識別不良信用風險，此節點的表現不佳。

選取節點中的觀察值

讓我們看一看節點 9 的觀察值，了解資料是否呈現任何其他有用的資訊。

- 連按兩下「瀏覽器」中的樹狀結構，開啟「樹狀結構編輯程式」。
- 按一下節點 9 以表示選取。(若要選取多個節點，可以按住 Ctrl 鍵來選擇)。
- 從「樹狀結構編輯程式」功能表選擇：
  規則 > 篩選觀察值...

圖表 4-16
「過濾觀察值」對話方塊



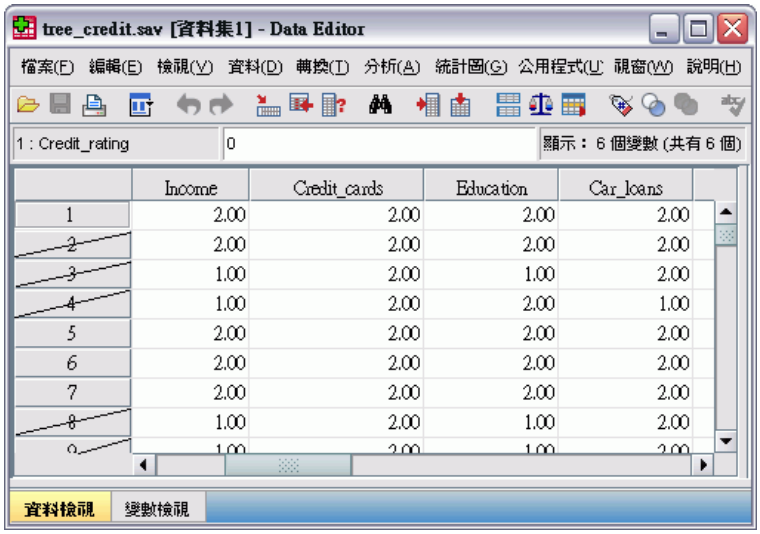
「過濾觀察值」對話方塊會建立過濾變數，並根據該變數的值來套用過濾設定。預設的過濾變數名稱為 filter_$。

- 所選節點的觀察值會得到數值為 1 的過濾變數。
- 所有其他觀察值會得到數值為 0 的過濾變數，而且這些觀察值會在後續的分析中排除，直到您變更過濾狀態為止。

在此例中，這表示非節點 9 中的觀察值現在會被過濾出 (但不會刪除)。

- 按一下「**確定**」建立過濾變數，並套用過濾條件。

圖表 4-17
「資料編輯程式」中的過濾觀察值



在「資料編輯程式」中，被過濾出的觀察值會以對角斜線穿過列號的方式表示。不在節點 9 的觀察值會被過濾出。 節點 9 的觀察值未被過濾出，所以後續的分析只會包括節點 9 的觀察值。

檢驗所選的觀察值

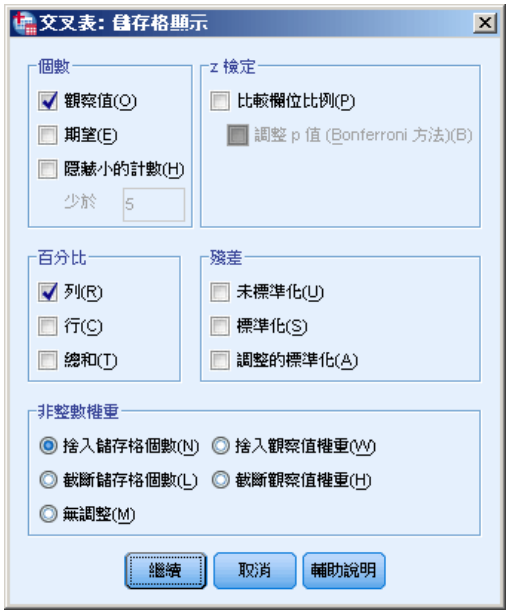
如同檢驗節點 9 中觀察值的第一個步驟，您要先看一下模式中沒有使用的變數。在此例中，分析時會包含資料檔案中的所有變數，但其中有兩個變數不會包含在最終模式中：教育和車貸。 因為可能沒有一個好的理由可以說明，為何程序要在最終模式中省略它們，或許便不多作解釋了，但我們無論如何還是再看一下。

- 從功能表選擇：
  分析(A) > 敘述統計 > 交叉表...

圖表 4-18
交叉表對話方塊

- 選取「信用評比」作為列變數。
- 選取「教育」和「車貸」作為行變數。
- 按一下「儲存格」。

圖表 4-19
交叉表儲存格顯示對話方塊



- 在「百分比」組別中，選取（勾選）「列」。

▶ 按一下「繼續」，再按一下主「交叉表」對話方塊中的「確定」以執行程序。

檢驗交叉表列，您可以看到對於模式中沒有的兩個變數，在優良和不良信用評比類別中的觀察值之間並沒有很大的差異。

圖表 4-20
所選節點內觀察值的交叉表列

Credit rating * Education 交叉表

			Education		
			High school	College	總和
Credit rating	Bad	個數	110	101	211
		Credit rating內的 %	52.1%	47.9%	100.0%
	Good	個數	128	144	272
		Credit rating內的 %	47.1%	52.9%	100.0%
總和		個數	238	245	483
		Credit rating內的 %	49.3%	50.7%	100.0%

Credit rating * Car loans 交叉表

			Car loans		
			None or 1	More than 2	總和
Credit rating	Bad	個數	18	193	211
		Credit rating內的 %	8.5%	91.5%	100.0%
	Good	個數	39	233	272
		Credit rating內的 %	14.3%	85.7%	100.0%
總和		個數	57	426	483
		Credit rating內的 %	11.8%	88.2%	100.0%

- 對於教育，略超過半數有不良信用評比的觀察值只有高中學教育程度，略超過半數有優良信用評比的觀察值具有大學教育程度 — 但此差異在統計上並不顯著。
- 對於車貸，只有一項或沒有車貸的優良信用觀察值百分比，高於不良信用觀察值的對等百分比，但這兩個組別中絕大部分的觀察值有二或多項車貸。

所以雖然您現在知道為何這些變數未包含在最終模式中，但很可惜的您並未對如何更佳的預測節點 9 有任何的洞察。如果該分析有其他的變數未指定，您可能要先檢驗這些變數，再繼續進行。

指定成本至結果

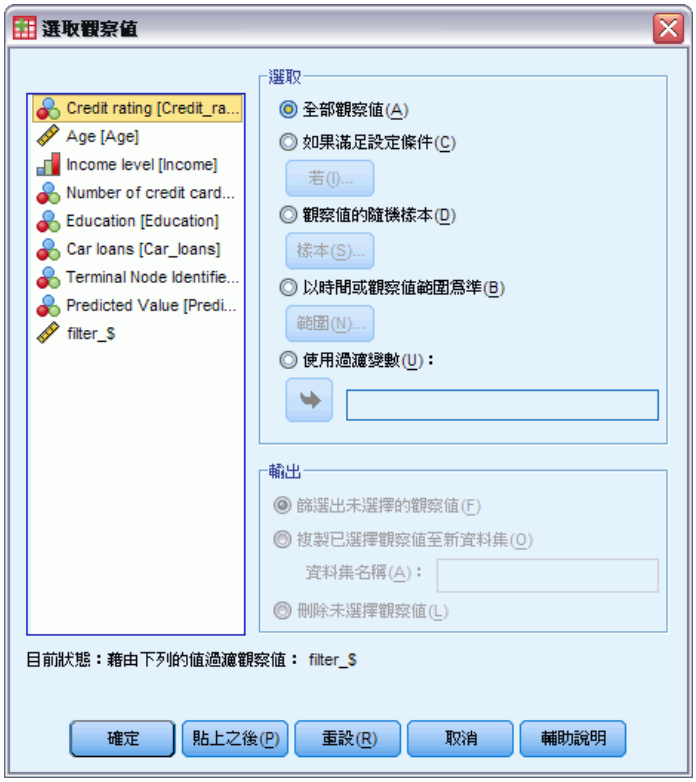
如同先前所提，除了節點 9 中幾乎一半的觀察值會落在每一個信用評比類別中之外，如果您主要的目標是建構一個可以正確識別不良信用風險的模式，那麼預測類別為「優良」會有問題。 雖然您可能無法改進節點 9 的表現，但您仍能將模式精確化，以改進不良信用評比觀察值的正確分類率 — 雖然這也會造成較高的優良信用評比觀察值的錯誤分類比率。

第一，您必須關閉觀察值過濾功能，讓所有的觀察值可再用於分析中。

▶ 從功能表選擇：
資料 > **選擇觀察值(S)...**

▶ 在「選取觀察值」對話方塊中，選取「**所有觀察值**」，再按一下「**確定**」。

圖表 4-21
「選擇觀察值」對話方塊



▶ 再次開啟「決策樹狀結構」對話方塊，按一下「**選項**」。

▶ 按一下「錯誤分類成本」索引標籤。

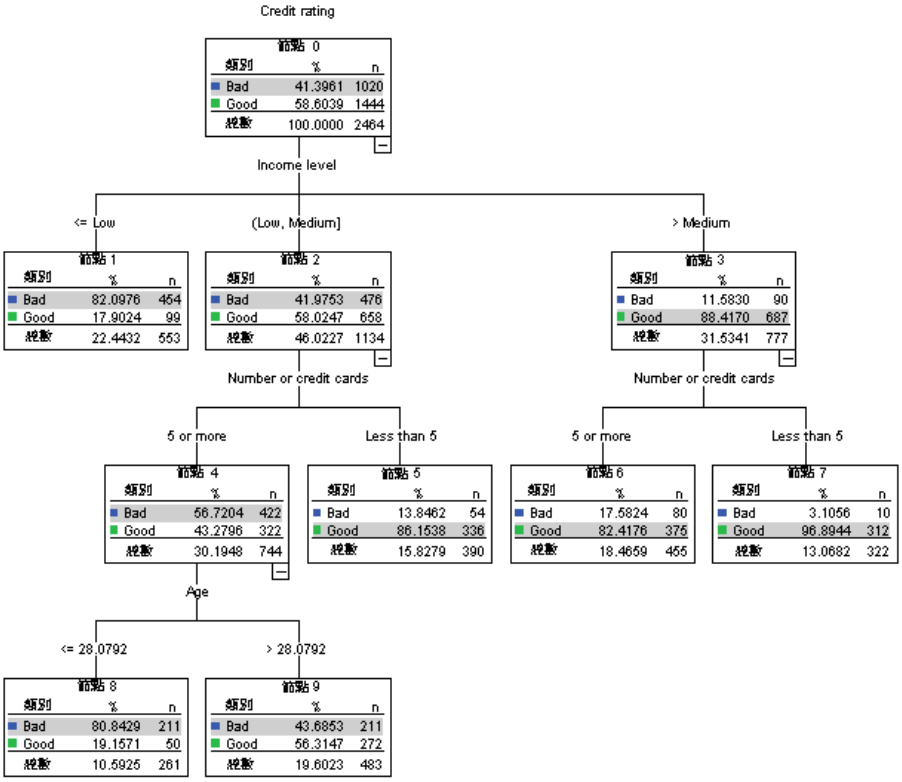
圖表 4-22
「選項」對話方塊，「錯誤分類成本」索引標籤

▶ 選取「自訂」，輸入 2 作為不良實際類別/ 優良 預測類別的數值。

這會告訴程序將不良信用風險錯誤分類成優良的「成本」，是將優良信用風險錯誤分類成不良「成本」的兩倍。

▶ 按一下「繼續」，再按一下主對話方塊中的「確定」來執行程序。

圖表 4-23
已調整成本值的樹狀模式



第一眼看時，程序產生的樹狀結構看起來與原始的樹狀結構大致相同。但是再仔細一點看，會發現雖然每個節點中觀察值的分散狀況沒有改變，但有些預測類別已改變。

對於終端節點，所有節點中的預測類別都維持相同，但以下節點除外：節點 9。預測類別現在為不良，即使略半數以上的觀察值是在優良類別中。

由於我們曾說過將不良信用風險錯誤分類為優良的成本較高，所以若有任何節點其觀察值平均分散在兩個類別之間，則這些節點的預測類別現在為不良，即使約有半數的觀察值是在優良類別中。

預測類別中的這個改變，會反映在分類表中。

圖表 4-24
根據已調整成本的風險和分類表

風險

估計	標準錯誤
.288	.011

成長方法: CHAID
依變數: Creditrating

分類

	預測的		
觀察的	Bad	Good	百分比修正
Bad	876	144	85.9%
Good	421	1023	70.8%
整體百分比	52.6%	47.4%	77.1%

成長方法: CHAID
依變數: Creditrating

- 幾乎 86% 的不良信用風險現在已正確的分類，相較起來，過去只有 65%。
- 另一方面，優良信用風險的正確分類已從 90% 降至 71%，整體正確分類已從 79.5% 降至 77.1%。

另請注意，風險估計和整體正確分類率不再彼此一致。如果整體正確分類率為 77.1%，那麼您可以預期風險估計為 0.229。 在此例中，增加錯誤分類不良信用觀值成本會讓風險值提高，讓它的解釋較不直覺。

摘要

您可以使用三個模式來將觀察值分類為由某些特性所識別的組別，例如與具有優良或不良信用記錄的銀行客戶有關之特性。如果某個預測結果比其他可能的結果來的重要，您可以將模式精確化，將該結果與較高的錯誤分類成本連結 — 但將一個結果的錯誤分類比率降低會增加其他結果的錯誤分類比率。

章節
5

建立評分模式

「決策樹狀結構」程序中最強大且最有用的一項功能就是建立模式的能力，所建立的模式可套用到其他資料檔以預測結果。例如，根據含有人口統計資訊及車輛購買價格資訊的資料檔，我們可以建立模式來預測人口統計特色類似的人可能會花多少錢購買新車—，再將這個模式套用到含有人口統計資訊但欠缺先前車輛購買資訊的其他資料檔。

在這個範例中，我們將使用資料檔 tree_car.sav。 如需詳細資訊，請參閱附錄 A 中的範例檔案中的IBM SPSS Decision Trees 20。

建立模式

- 若要執行「決策樹狀結構」分析，請從功能表選擇：
  分析(A) > 分類 > 樹...

圖表 5-1
「決策樹狀結構」對話方塊

- 選取「主要車輛價格」作為依變數。
- 選取所有其他的變數作為自變數。(此程序會自動排除任何對最終模式沒有顯著貢獻的變數)。
- 請選擇「CRT」作為成長方法。

Copyright IBM Corporation 1989, 2011.

- 按一下「輸出」。

圖表 5-2
「輸出」對話方塊，「規則」索引標籤

- 按一下「**規則**」索引標籤。
- 選取（勾選）「**產生分類規則**」。
- 選取 **IBM SPSS Statistics** 作為語法。
- 選取「**指派觀察值的數值**」作為「類型」。
- 選取（勾選）「**匯出規則到檔案**」並輸入檔名及目錄位置。

  請記住或寫下檔名及位置，因為您稍後會需要這些資訊。如果您沒有記下目錄的路徑，您可能會不知道檔案儲存的位置。您可以利用「**瀏覽**」按鈕來瀏覽到特定（且有效）的目錄位置。

- 按一下「**繼續**」，再按一下「**確定**」，執行程序並建立樹狀結構。

評估模式

在套用模式到其他資料檔之前，您可能會希望確定模式可和建立模式的原始資料充分搭配。

模式摘要

圖表 5-3
模式摘要表

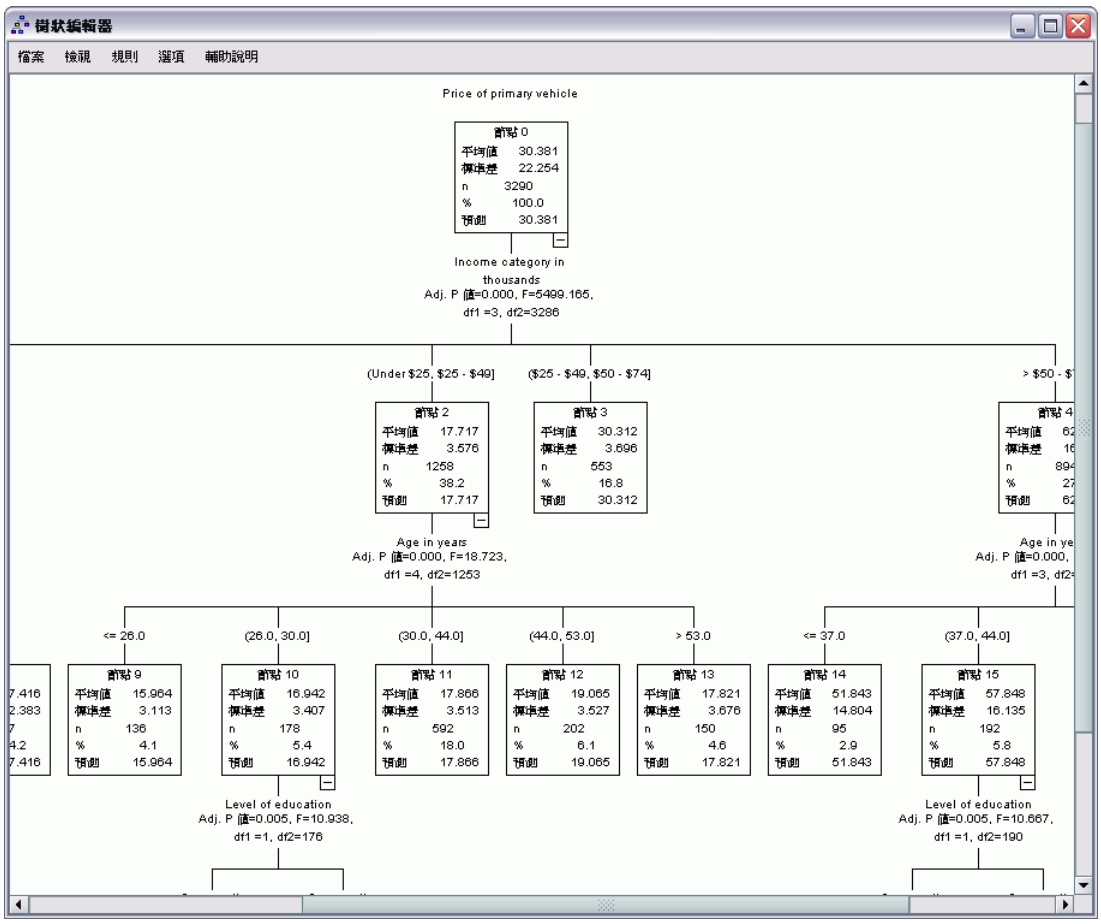
規格	成長方法	CRT
	依變數	Price of primary vehicle
	自變數	Age in years, Gender, Income category in thousands, Level of education, Marital status
	有效性	無
	最大樹狀結構深度	5
	父節點中最小的觀察值	100
	子節點中最小的觀察值	50
結果	所包含的自變數	Income category in thousands, Age in years, Level of education, Marital status
	節點數量	29
	終端節點數量	15
	深度	5

模式摘要表指出，在選定的自變數中，只有三個變數有顯著貢獻，足以列入最終模式中：收入、年齡和教育程度。如果您想要將這個模式套用到其他資料檔，由於模式中使用的自變數必須出現在您想要套用模式的任何資料檔，因此您必須了解這項資訊。

摘要表也指出，由於樹狀結構模式有 29 個節點及 15 個終點，因此模式本身可能不是特別簡單。如果您想要的是可實際套用的可靠模式，而不是容易描述或解釋的簡單模式，這可能不會是問題。當然，就實際情況而言，您可能也想要一個不需依賴太多自變數（預測值）的模式。在這個情況下，由於只有三個自變數列入最終模式，這不會是問題。

樹狀結構模式圖

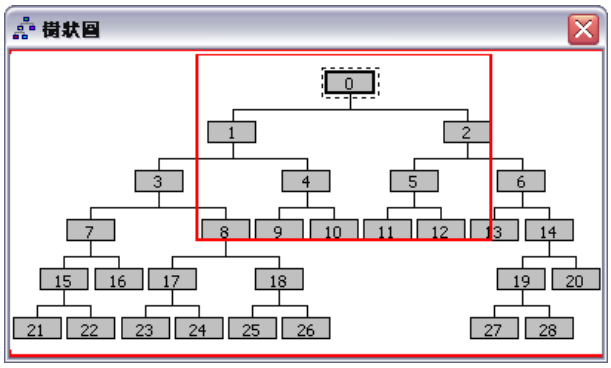
圖表 5-4
「樹狀結構編輯程式」中的「樹狀結構」模式圖表



樹狀結構模式圖的節點很多，所以即使圖表大小仍足以讓您讀取節點內容資訊，您可能很難立刻看到整個模式。您可利用樹狀結構圖查看整個樹狀結構：

- 連按兩下「瀏覽器」中的樹狀結構，開啟「樹狀結構編輯程式」。
- 從「樹狀結構編輯程式」功能表選擇：
  檢視 > 樹狀圖

圖表 5-5
樹狀結構圖。



- 樹狀結構圖顯示整個樹狀結構。您可以變更樹狀結構圖視窗的大小，然後該視窗會放大或縮小樹狀結構的地圖顯示，以配合視窗大小。
- 樹狀結構圖中的反白區域就是目前顯示於「樹狀結構編輯程式」中的樹狀結構區域。
- 您可利用樹狀結構圖瀏覽樹狀結構並選取節點。

如需詳細資訊，請參閱第 36 頁第 2 章中的樹狀圖。

至於尺度依變數，每個節點都顯示依變數的平均數及標準差。節點 0 顯示整體平均車輛購買價格約為 29.9（以千元為單位），標準差約為 21.6。

- 節點 1，代表在收入少於 75（以千元為單位）的觀察值中，平均車輛購買價格只有 18.7。
- 相反的，節點 2代表在收入等於或高於 75（以千元為單位）的觀察值中，平均車輛購買價格為 60.9。

更進一步的樹狀結構調查會顯示年齡及教育程度也與車輛購買價格有關，但現在我們主要是對模式的實際應用有興趣，而不是其成分的詳細檢驗。

風險估計

圖表 5-6
風險表

風險

估計	標準錯誤
68.485	2.985

成長方法: CRT
依變數: Price of primary vehicle

我們檢驗的結果並沒有告訴我們這是不是一個特別好的模式。模式效能的指標之一是風險估計。對於尺度依變數而言，風險估計就是節點內變異數的量數，這個量數本身可能不會告訴您太多資訊。較低的變異數代表比較好的模式，但是變異數與量數單位是相關的。例如，如果價格是以元而不是以千元記錄，則風險估計將成長為一千倍。

若要為含有尺度依變數的風險估計提供有意義的解讀，必須先進行幾項工作：

- 變異數總數等於節點內（誤差）變異數加上節點間（已說明的）變異數。
- 節點內變異數是風險估計值： 68.485.
- 變異數總數是考量任何自變數前，依變數的變異數，也就是根節點的變異數。
- 根節點所顯示的標準差為 21.576；所以變異數總數就是這個數值的平方： 465.524.
- 因為誤差而造成的變異數（未說明的變異數）比例為 68.485 / 465.524 = 0.147。
- 由模式說明的變異數比例為 1 - 0.147 = 0.853，或 85.3%，顯示這是很好的模式。（對於類別依變數而言，這個解讀類似整體正確分類比率）。

套用模式到另一個資料檔

判定這個模式很好之後，我們現在可以將模式套用到包含類似的年齡、收入，及教育程度等變數的其他資料檔，並產生代表該檔案中每個觀察值中預測車輛購買價格的新變數。 這個程序通常稱為**評分**。

我們產生模式時，已經指定指派觀察值數值的「規則」應該儲存在文字檔中，以指令語法的形式。我們現在將使用該檔案中的指令在另一個資料檔中產生分數。

▶ 開啟資料檔 tree_score_car.sav。如需詳細資訊，請參閱附錄 A 中的範例檔案中的 IBM SPSS Decision Trees 20。

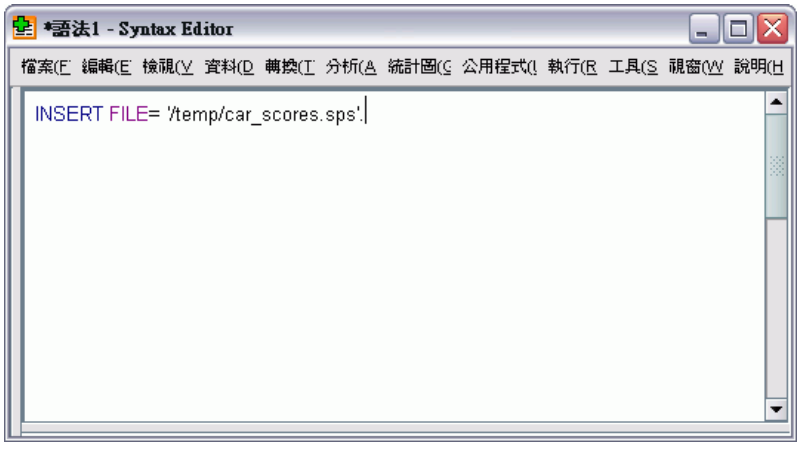
▶ 下一步，從功能表選擇：
檔案 > 開啟新檔(N) > 語法

▶ 在指令語法視窗中輸入：

```
INSERT FILE=
 '/temp/car_scores.sps'.
```

如果您曾使用不同的檔名或位置，請進行適當變更。

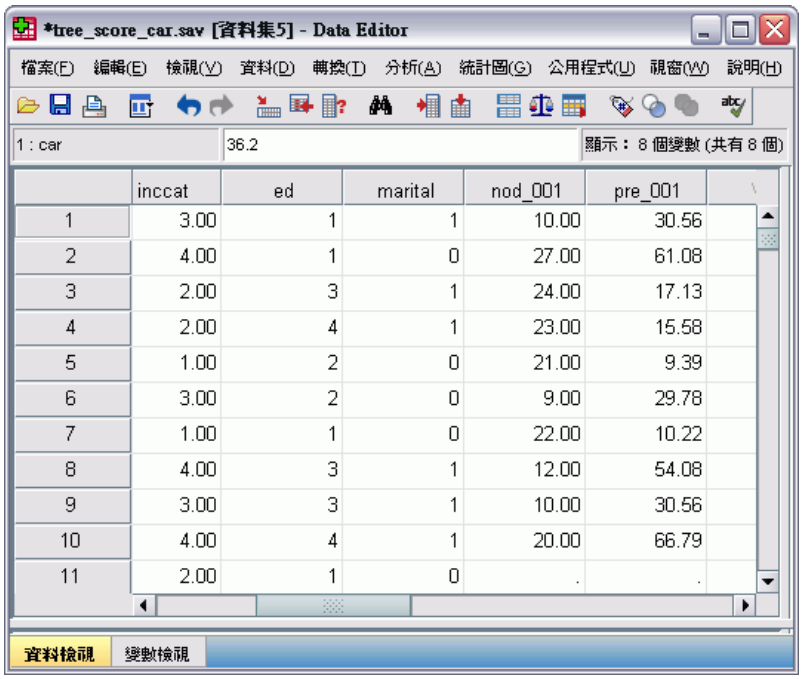
圖表 5-7
以「INSERT」指令執行指令檔的語法視窗



INSERT 指令將在指定檔案內執行指令，也就是我們建立模式時所產生的「規則」檔案。

▶ 從指令語法視窗功能表選擇：
執行(R) > 全部(A)

圖表 5-8
預測值已新增到資料檔



	inccat	ed	marital	nod_001	pre_001
1	3.00	1	1	10.00	30.56
2	4.00	1	0	27.00	61.08
3	2.00	3	1	24.00	17.13
4	2.00	4	1	23.00	15.58
5	1.00	2	0	21.00	9.39
6	3.00	2	0	9.00	29.78
7	1.00	1	0	22.00	10.22
8	4.00	3	1	12.00	54.08
9	3.00	3	1	10.00	30.56
10	4.00	4	1	20.00	66.79
11	2.00	1	0	.	.

兩個變數會新增到資料檔：

- nod_001 包含模式為每個觀察值預測的終端節點數量。
- pre_001 包含每個觀察值中車輛購買價格的預測值。

由於我們曾要求指派終點數值的規則，可能的預測值數量等於終點數量，在這個範例中為 15。例如，預測節點數量為 10 的每個觀察值將擁有相同的預測車輛購買價格：30.56. 這也是原始模式中終端節點 10 的平均數值，並非巧合。

雖然您一般會將模式套用到依變數值未知的資料，但是在這個範例中，我們實際套用模式的資料檔已經包含了依變數的數值，—您可以比較模式預測值與實際值。

▶ 從功能表選擇：
分析(A) > 相關 > 雙變數...

▶ 選取「主要車輛價格」及「pre_001」。

圖表 5-9
「雙變數相關分析」對話方塊

▶ 按一下「確定」執行程序。

圖表 5-10
實際及預測車輛價格的相關

		Price of primary vehicle	pre_001
Price of primary vehicle	Pearson 相關	1	.919**
	顯著性(雙尾)		.000
	個數	3290	3290
pre_001	Pearson 相關	.919**	1
	顯著性(雙尾)	.000	
	個數	3290	3290

**. 在顯著水準為0.01時(雙尾)，相關顯著。

0.92 的相關指出實際及預測車輛價格之間有很高的正相關，代表模式很有用。

摘要

您可以使用「決策樹狀結構」程序建立模式，再將模式套用到其他資料檔以預測結果。目標資料檔必須包含名稱與最終模式中自變數名稱相同的變數，以相同的計量單位測量且包含相同的使用者定義的遺漏值（如果有的話）。但是，最終模式排除的依變數或自變數都不需出現在目標資料檔。

章節
6

樹狀結構模式中的遺漏值

不同的成長方法可以不同的方式處理自變數（預測值）的遺漏值：

- CHAID 和 Exhaustive CHAID 會將每個自變數的所有系統遺漏值和使用者遺漏值視為單一類別對於尺度和次序自變數，該類別隨後是否會與自變數的其他類別合併，視成長條件而定。
- CRT 和 QUEST 會嘗試使用自變數（預測值）的**代理**。對於該變數值已遺漏的觀察值而言，會使用其他具有與原始變數高度關聯的自變數來進行分類。這些替代的預測值稱為代理。

此例顯示當模式中使用的自變數有遺漏值時，CHAID 和 CRT 之間的差異。

在這個範例中，我們將使用資料檔 tree_missing_data.sav。如需詳細資訊，請參閱附錄 A 中的範例檔案中的IBM SPSS Decision Trees 20。

注意：對於名義自變數和名義依變數，您可以選擇將**使用者遺漏**值作為有效值，在此狀況下這些值被視為任何其他未遺漏值。 如需詳細資訊，請參閱第 20 頁第 1 章中的遺漏值。

以 CHAID 分類的遺漏值

圖表 6-1
將資料記為遺漏值

Copyright IBM Corporation 1989, 2011.

和信用風險範例一樣，(如需詳細資訊，請參閱第 4 章)，此例也會嘗試建構出一個模式，將優良信用風險和不良信用風險分類。主要差異是此資料檔案包含模式中所使用某些自變數的遺漏值。

- 若要執行「決策樹狀結構」分析，請從功能表選擇：
  分析(A) > 分類 > 樹...

圖表 6-2
「決策樹狀結構」對話方塊

- 選取「信用評比」作為依變數。
- 選取所有其它的變數作為自變數。(此程序會自動排除任何對最終模式沒有顯著貢獻的變數)。
- 對於成長方法，請選取「CHAID」。

  就此例而言，我們要保持樹狀結構十分的簡單，所以我們要增加父節點和子節點觀察值的最小值，來限制樹狀結構的成長。

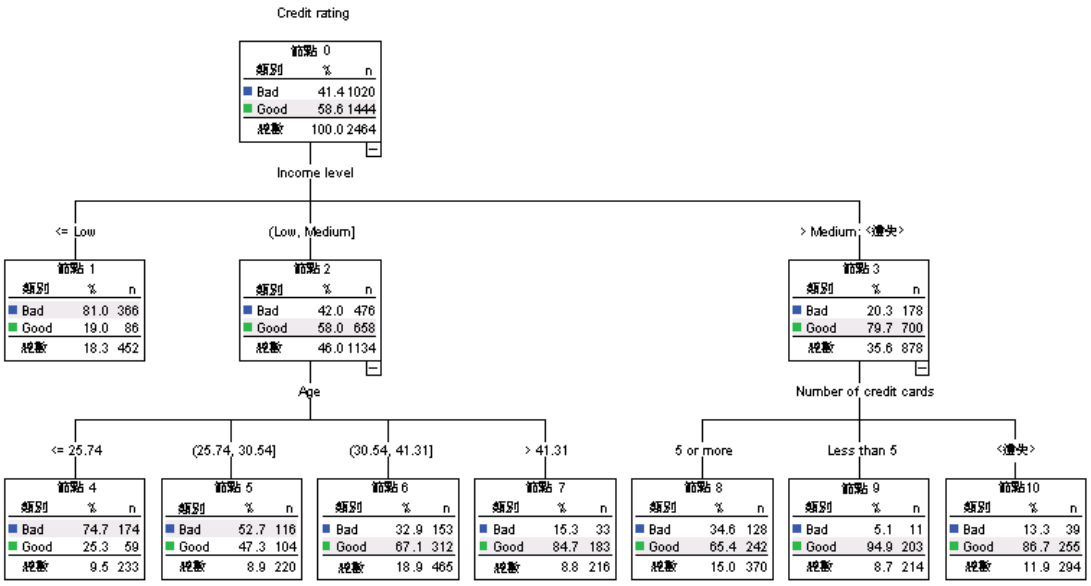
- 在主「決策樹狀結構」對話方塊中，按一下「條件」。

圖表 6-3
「條件」對話方塊，「成長限制」索引標籤

- 對於「最小觀察值數」，輸入 400 作為「父節點」，輸入 200 作為「子節點」。
- 按一下「繼續」，再按一下「確定」來執行程序。

CHAID 結果

圖表 6-4
含遺漏自變數值的 CHAID 樹狀結構



對於節點 3，收入水準會顯示為 >Medium;<missing>。這表示節點包含高收入類別的觀察值，加上任何具收入水準遺漏值的觀察值。

終端節點 10 包含具有信用卡數目遺漏值的觀察值。如果您對於識別優良信用風險有興趣，這實際上是次佳的終端節點；如果您要使用此模式來預測優良信用風險，可能發生問題。您或許不需要一個預測信用評比的模式，只因為您不知道一個觀察值有多少張信用卡，這些觀察值中有一些可能也遺漏收入水準資訊。

圖表 6-5
CHAID 模式的風險和分類表

風險

估計	標準錯誤
.249	.009

成長方法: CHAID
依變數: Creditrating

分類

	預測的		
觀察的	Bad	Good	百分比修正
Bad	656	364	64.3%
Good	249	1195	82.8%
整體百分比	36.7%	63.3%	75.1%

成長方法: CHAID
依變數: Creditrating

風險和分類表指示 CHAID 模式正確分類約 75% 的觀察值。這並不差，但也不好。此外，我們有理由懷疑優良信用觀察值的正確分類率可能過度樂觀，因為它有一部分所依據的假設是所缺少的兩個自變數（收入水準和信用卡數）相關資訊是優良信用的指標。

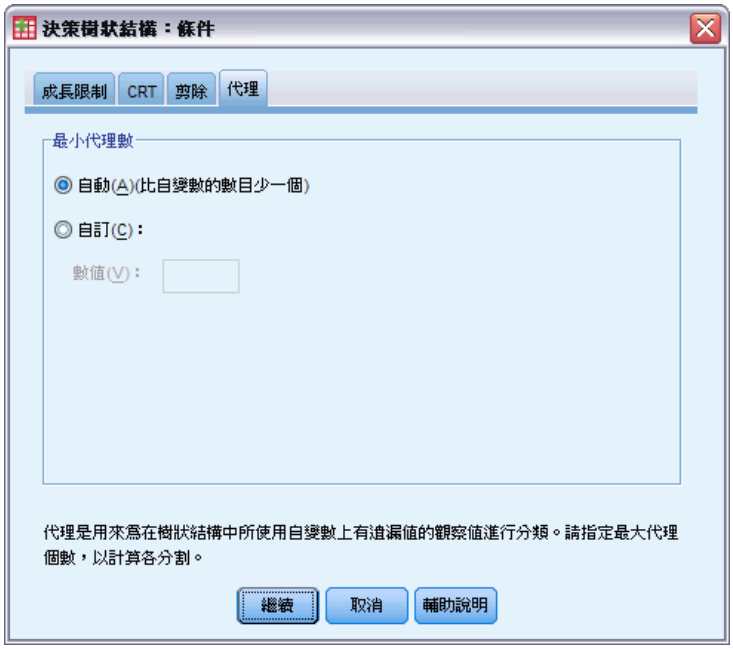
以 CRT 分類的遺漏值

現在讓我們嘗試相同的基本分析，但這次我們使用 CRT 作為成長方法。

- 在主要「決策樹狀結構」對話方塊中，選取「CRT」作為成長方法。
- 按一下「條件」。
- 確定父節點的觀察值最小數仍設為 400，子節點為 200。
- 按一下「代理」索引標籤。

注意：除非您選取「CRT」或「QUEST」作為成長方法，否則不會看到「代理」索引標籤。

圖表 6-6
「條件」對話方塊，「代理」索引標籤



對於每個自變數節點分割，**自動**設定會視每個為模式指定的其他自變數為可能的代理。由於在此例中沒有很多的自變數，因此**自動**設定是適合的。

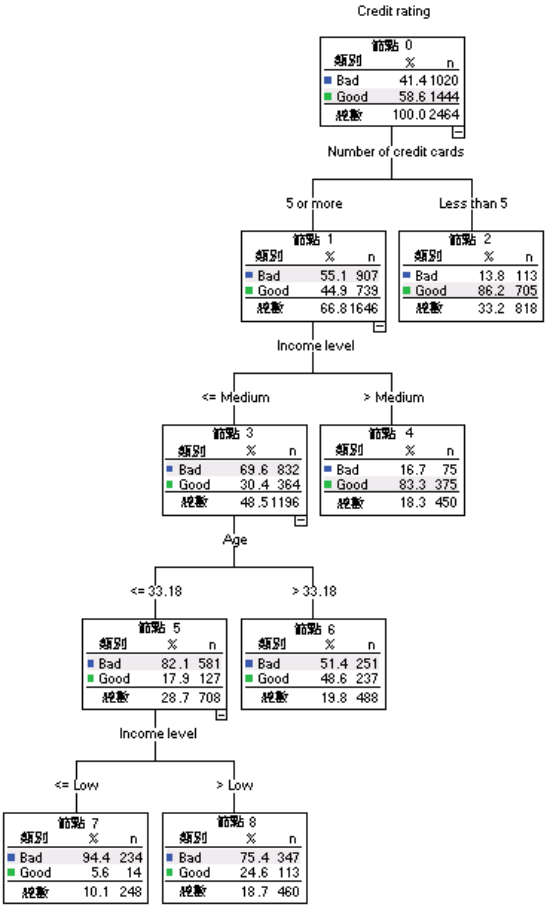
- 按一下「**繼續**」。
- 在主「決策樹狀結構」對話方塊中，按一下「**輸出**」。

圖表 6-7
「輸出」對話方塊，「統計量」索引標籤

- 按一下「統計量」索引標籤。
- 選取「根據分割來代理」。
- 按一下「繼續」，再按一下「確定」來執行程序。

CRT 結果

圖表 6-8
含遺漏自變數值的 CRT 樹狀結構



您可能馬上就注意到這個樹狀結構看起來和 CHAID 樹狀結構差很多。但它本身不一定具有很大意義。在 CRT 樹狀結構模式中，所有的分割都是二元的，也就是每個父節點只會分成兩個子節點。在 CHAID 模式中，父節點可分成許多子節點。所以樹狀結構通常看起來會不一樣，即使它們都代表相同的基礎模式。

但是，仍有一些重要的差異：

- 在 CRT 模式中最重要的自變數（預測值）是信用卡數，而在 CHAID 模式中，最重要的預測值是收入水準。
- 對於擁有不超過 5 張信用卡的觀察值，信用卡數是信用評比唯一顯著的預測值，節點 2 是終止節點。
- 使用 CHAID 模式時，收入水準和年齡也會包括在模式中，雖然收入水準現在是第二預測值，而非第一預測值。
- 沒有任何節點包含在 <missing> 類別中，因為 CRT 使用代理預測值，而非模式中的遺漏值。

圖表 6-9
CRT 模式的風險和分類表

風險

估計	標準錯誤
.226	.008

成長方法: CRT
依變數: Creditrating

分類

觀察的	預測的		
	Bad	Good	百分比修正
Bad	721	299	70.7%
Good	258	1186	82.1%
整體百分比	39.7%	60.3%	77.4%

成長方法: CRT
依變數: Creditrating

- 風險和類別表顯示整體的正確分類幾乎為 78%，較 CHAID 模式 (75%) 略為增加。
- CRT 模式中不良信用觀察值的正確分類率 —81.6%，與 CHAID 模式中只有 64.3% 相較起來高出許多。
- 但是，優良信用觀察值的正確分類率已由 CHAID 的 82.8%，下降至 CRT 的 74.8%。

代理

CHAID 和 CRT 模式之間的差異有一部分是導因於使用 CRT 模式中的代理。代理表顯示代理在模式中的使用方式。

圖表 6-10
代理表

父節點	自變數		增進	關聯
0	主要的	Number of credit cards	.090	
	代理人	Car loans	.052	.643
		Age	.001	.004
1	主要的	Income level	.071	
	代理	Age	.001	.004
3	主要的	Age	.022	
7	主要的	Income level	.006	
	代理	Age	.000	.009

成長方法: CRT
依變數: Credit rating

- 在根節點（節點 0）處，最佳自變數（預測值）是信用卡數。
- 對於任何有信用卡數遺漏值的觀察值，會使用車貸作為代理預測值，因為此變數與信用卡數有高度相關 (0.643)。
- 如果觀察值也有車貸遺漏值，再使用年齡作為代理（雖然它的相關值十分低，只有 0.004)。
- 年齡也作為節點 1 和 5 中收入水準的代理。

摘要

不同的成長方法以不同的方式處理遺漏值。如果用來建立模式的資料包含許多遺漏值—，或者您要將該模式套用在任何包含遺漏值的其他資料檔案上—，則您應該評估各種模式的遺漏值作用。如果您要使用模式中的代理來補償遺漏值，可使用 CRT 或 QUEST 方法。

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。
- **anorectic.sav。** 在將厭食/暴食行為症狀學標準化的過程中，研究人員 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。
- **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 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。

Copyright IBM Corporation 1989, 2011.

- **breakfast.sav。**在經典研究中，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 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 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。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。
- **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」 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 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 以此表說明可能的社會團體。Guttman 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 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，、 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。
- **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。
- **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 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。
- **stocks.sav** 本假設資料檔包含一年的股票價格和數量。
- **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 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 出現和分析過。
- **verd1985.sav。** 本資料檔有關於一項調查 。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 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。
- **worldsales.sav** 本假設資料檔包含依洲和產品分類之銷貨收益。

附錄
B

注意事項

本資訊適用於全球提供之產品與服務。

IBM 可能並未在其他國家提供在本文件中討論到的產品、服務或功能。有關目前在貴地區可供使用的產品與服務相關資訊，請洽您當地的 IBM 服務代表。對於 IBM 產品、程式或服務的任何參考，目的並不是要陳述或暗示只能使用 IBM 產品、程式或服務。任何功能相等且未侵犯 IBM 智慧財產權的產品、程式或服務皆可使用。但是，評估及確認任何非 IBM 產品、程式或服務的操作之責任應由使用者承擔。

IBM 可能有一些擁有專利或專利申請中的項目包含本文件所描述的內容。本文件的提供並不表示授與您對於這些專利的權利。您可以將書面的授權查詢寄至：

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

對於與雙位元組字元集 (DBCS) 資訊相關的授權查詢，請與貴國的 IBM 智慧財產部門聯絡，或將查詢郵寄至：

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

下列條款不適用於英國，或其他任何當地法律規定與下列條款不一致的國家：
INTERNATIONAL BUSINESS MACHINES 聲明係以「現狀」提供，沒有任何保固；不作任何明示或默示的保證，包括但不限於不侵權、適銷性或適合某一特定用途之保證。某些州不允許特定交易中明示或默示的保固聲明，因此，此聲明或許對您不適用。

此資訊內容可能包含技術失準或排版印刷錯誤。此處資訊會定期變更，這些變更將會納入新版的聲明中。IBM 可能會隨時改善和 / 或變更此聲明中所述的產品和 / 或程式，恕不另行通知。

此資訊中對於非 IBM 網站之任何參考僅為查閱方便而設，而且在任何情況中均不得作為那些網站之背書。該「網站」的內容並非此 IBM 產品的部分內容，使用該「網站」需自行承擔風險。

IBM 可能會以任何其認為適當的方式使用或散佈您提供的任何資訊，無需對您負責。

意欲針對達成以下目的而擁有本程式相關資訊之程式被授權人：(i) 在獨立建立的程式與其他程式 (包括本程式) 之間交換資訊及 (ii) 共用已交換的資訊，應聯絡：

IBM Software Group，收件人：Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA。

在適當條款與條件之下，包括某些情況下 (支付費用)，或可使用此類資訊。

在本文件中描述的授權程式及其適用之所有授權材料皆由 IBM 在與我方簽訂之 IBM 客戶合約、IBM 國際程式授權合約或任何相等效力合約中提供。

有關非 IBM 產品的資訊來自於那些產品的供應商、其公佈內容或其他可公開取得的來源。IBM 尚未測試那些產品，且無法確認效能準確度、相容性或任何其他與非 IBM 產品相關的理賠。對於非 IBM 產品之功能有任何問題，應由那些產品之供應商加以解決。

Copyright IBM Corporation 1989, 2011.

此資訊包含用於日常企業運作的資料和報表範例。為了儘可能提供完整說明，範例中包含了人名、公司名稱、品牌名稱和產品名稱。所有的名稱皆為虛構，使用的名稱或地址和實際的企業如有雷同純屬巧合。

如果您正在螢幕上檢視此資訊，則圖片和彩色說明可能不會顯示。

商標

IBM、IBM 標誌、ibm.com 和 SPSS 為 IBM Corporation 之註冊商標，已經於世界各地許多法律管轄區域註冊。IBM 註冊商標的清單目前可於「網站」上取得，網址為：http://www.ibm.com/legal/copytrade.shtml。

Adobe、Adobe 標誌、PostScript 以及 PostScript 標誌為 Adobe Systems Incorporated 於美國和 / 或其他國家的註冊商標或商標。

Intel、Intel 標誌、Intel Inside、Intel Inside 標誌、Intel Centrino、Intel Centrino 標誌、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 為 Intel Corporation 或其分公司於美國和其他國家的商標或註冊商標。

Java 和所有 Java 商標和標誌皆為 Sun Microsystems, Inc. 於美國和 / 或其他國家的商標。

Linux 為 Linus Torvalds 於美國和 / 或其他國家的註冊商標。

Microsoft、Windows、Windows NT 和 Windows 標誌為 Microsoft Corporation 於美國和 / 或其他國家的商標。

UNIX 為 The Open Group 於美國和其他國家的註冊商標。

本產品使用 WinWrap Basic，Copyright 1993-2007，Polar Engineering and Consulting，http://www.winwrap.com。

其他產品和服務名稱可能為 IBM 或其他公司的商標。

Adobe 產品的擷取畫面已取得 Adobe Systems Incorporated 之翻印許可。

Microsoft 產品的擷取畫面已取得 Microsoft Corporation 之翻印許可。

索引