IBM SPSS Data Preparation 29

附註

使用此資訊和支援的產品之前，請先閱讀第 19 頁的『注意事項』中的資訊。

產品資訊

除非新版中另有指示，否則此版本適用於 IBM® SPSS Statistics 29 版次 0，修訂 1 版及所有後續版次與修訂。

目錄

第 1 章 資料準備

Base Edition 中包含下列資料準備功能。

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」容許您識別作用中的資料集中的異常觀察值和無效觀察值、變數和資料值，並準備建模資料。

「資料準備」程序的用法

「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：

- **meta 資料準備**。 檢視您資料檔中的變數並決定其有效數值、標記及測量層級。 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 meta 資料準備中的另一個步驟。
- **模型準備**。 使用自動資料準備以取得可改善模型建置的原始欄位轉換。 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 meta 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程式模組建置模型。

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表示式定義。

驗證規則會儲存在您資料檔的資料目錄。 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

您可從安裝隨附的外部資料檔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

1. 在功能表上，選擇：

   資料 > 驗證 > 載入預先定義的規則...

您可改用「複製資料內容精靈」，從任何資料檔載入規則。

定義驗證規則

「定義驗證規則」對話框可讓您建立並檢視單一變數與交叉變數驗證規則。

建立和檢視驗證規則

1. 從功能表中選擇：

   資料 > 驗證 > 定義規則...

此對話框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。

2. 在「單一變數規則」和「交叉變數規則」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

「單一變數規則」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。 開啟對話框時，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：

- **新建**。 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

  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。

  若要定義範圍，請輸入最小值或最大值，或兩者。 勾選框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

  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。

  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，勾選框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

「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。 開啟對話框時，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：

- **新建**。 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」，依此類推。
- **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。
- **邏輯表示式**。 事實上，這是規則定義。 您應編碼表示式，將無效觀察值評估為 1。

建立表示式

1. 若要建立表示式，請將組成成份貼入「表示式」欄位，或者直接輸入「表示式」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**)。 輸入標有問號的所有參數的值（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。 對話框的保留區域會顯示簡要的說明，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。
- 如果值包含小數，必須使用句點 (.) 作為小數指示符。

驗證資料

「驗證資料」對話框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」對話框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

1. 從功能表中選擇:

   資料 > 驗證 > 驗證資料...

2. 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:

3. 按一下「**交叉變數規則**」標籤，並且套用一個或多個交叉變數規則。

視需要而定，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）觀察值 ID 變數時，會將這些數值的組合當作觀察值 ID 來處理。

具有未知測量層級的欄位(F)

若在資料集中出現一或多個未知的變數 (欄位) 測量層級，就會顯示「測量層級」警示。 由於測量層級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層級。

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層級的任何欄位指派預設的測量層級。 若為大型資料集，則讀取時可能需要一些時間。

手動指派。 開啟對話框，以列出具有未知測量層級所有欄位。 您可以使用此對話框以指派測量層級給這些欄位。 您可以在「資料編輯器」的「變數視圖」中指派測量層級。

由於測量層級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層級之前，無法存取對話框來執行此程序。

驗證資料基本檢查

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

分析變數。 如果已經選擇「變數」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。 勾選框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。 指定值必需是小於或等於 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 ID。 如果已經選擇「變數」標籤上的任何個案 ID 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個案 ID 的個案。 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 ID 為不完整。
- **標示重複 ID**。 這個選項會報告含有重複觀察值 ID 的觀察值。 會將不完整 ID 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」標籤上所定義的分析變數。

驗證資料單一變數規則：

「單一變數規則」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**」。 如需相關資訊，請參閱主題第 2 頁的『定義單一變數規則』。

分析變數。 這個清單會顯示分析變數、摘要其分佈狀態，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。

變數分佈。 「分析變數」清單中的所顯示分佈摘要，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」文字方框中所指定。 按一下「**重新掃描**」，更新分佈摘要。

驗證資料交叉變數規則

「交叉變數規則」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**」。 如需相關資訊，請參閱主題第 2 頁的『定義交叉變數規則』。

驗證資料輸出

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。

顯示分析變數的敘述性統計量。 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。

將違反驗證規則的觀察值移動到作用中資料集的上方。 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

「儲存」標籤可以讓您儲存將規則違規記錄到作用中資料集的變數。

彙總變數。 這些是可以儲存的個別變數。 核取一個方框以儲存變數。 會提供變數的預設名稱，您可以進行編輯。

- **空白觀察值指標**。 空的觀察值會獲指派值 1。 所有其他觀察值都已編碼 0。 變數值反映在基本檢查標籤上指定的範圍。
- **重複 ID 群組**。 含有相同觀察值 ID 的觀察值（而不是含有不完整 ID 的觀察值），會指定相同的組別號碼。 會將含有唯一或不完整 ID 的觀察值編碼為 0。
- **不完整的 ID 指標**。 具有空值或不完整案例 ID 的觀察值，會指派值 1。 所有其他觀察值都已編碼 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。

取代現有的彙總變數。 儲存於資料檔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。

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置模式，不需事先瞭解統計相關概念。 模式將可更快地建置並進行資料評分，此外，使用 ADP 可提高自動建模程序。

附註：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。舊欄位未用於進一步分析；其角色設定為「無」。 另請注意，任何使用者遺漏值資訊都不會傳送至這些新建立的欄位，而新欄位中的任何遺漏值都是系統遺漏的。

範例。 某資源有限的保險公司，打算調查屋主的保險理賠，希望建置標示可疑潛在詐欺理賠的模型。 建置模型之前，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，因此會在互動式模型中使用自動資料準備。

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車型，他們希望建立汽車銷售額與汽車特性之間的關係。 他們會使用自動資料準備來準備分析用的資料，並使用準備「之前」與「之後」的資料建置模型，以瞭解結果有何差異。

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置模型的速度並提升這些模型的預測能力。 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應遵循的模型建置優先順序。

- **權衡速度與準確度**。 此選項準備資料時，會兼顧模型建置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，會優先考慮模型建置演算法處理資料的速度。 當您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，會優先考慮模型建置演算法所產生預測的準確度。
- **自訂分析**。 當您想在「設定」標籤中手動變更演算法時，請選取此選項。 請注意，如果您之後變更了「設定」標籤上的選項，但該變更與其他目標之一不相容時，系統會自動選取此設定。

取得自動資料準備

在功能表上，選擇：

1. 在功能表上，選擇：

   轉換 > 準備資料進行建模 > 自動...

2. 按一下**執行**。

視需要而定，您可以：

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

取得互動式資料準備

1. 在功能表上，選擇：

   轉換 > 準備資料進行建模 > 互動式...
2. 在對話框上方的工具列中，按一下「**分析**」。
3. 按一下「分析」標籤並檢視建議的資料準備步驟。
4. 如果滿足您的需求，請按一下「**執行**」。 否則，請按一下「**清除分析**」，變更您要的任何設定，然後按一下「**分析**」。

視需要而定，您可以：

- 在「目標」標籤上指定目標。
- 在「欄位」標籤上指定欄位指派。
- 在「設定」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。

欄位標籤

「欄位」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，則會將其當作目標；否則將不會有目標。 含有預先定義「輸入」角色的所有欄位都將作為輸入。 至少需要一個輸入欄位。主題，以取得更多資訊。

使用自訂欄位指派。 從欄位的預設清單移動欄位來改寫欄位角色時，對話框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）**。 若您計劃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」類似。

設定標籤

「設定」標籤包含幾組不同的設定，您可以修改這些設定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」標籤會自動更新為選取「**自訂分析**」選項。

準備日期與時間

許多建模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，以用作您現有資料中日期和時間的模型輸入。 必須以日期或時間儲存類型預先定義包含日期與時間的欄位。 原始的日期與時間欄位在自動資料準備之後將不建議作為模型輸入。

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，作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，作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。

擷取循環時間元素。 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勾選框，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「**欄位名稱**」畫面中定義的字尾，並且會忽略原始日期。

- **從日期擷取**。 對於任何日期輸入，指定您要擷取年、月、日或任何組合。
- **從時間擷取**。 對於任何時間輸入，指定您要擷取小時、分鐘、秒或任何組合。

排除欄位

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功能等級品質。 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。 預設值為 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。 預設值為 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序數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，常數欄位還是會遭到自動排除。 預設值是 95。

調整測量

調整測量層級。 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。

測量層級。 指定含有「太少」值之連續欄位的測量層級是否可調整為序數，以及含有「太多」值之序數欄位的測量層級是否可調整為連續。

- **序數欄位數值的最大數量**。 超過指定類別數目的序數欄位會重新分配為連續欄位。 指定一個正整數。 預設值為 10。 此值必須大於或等於連續欄位值的最小數目。
- **連續欄位數值的最小數量**。 少於指定唯一值數目的連續欄位會重新分配為序數欄位。 指定一個正整數。 預設值為 5。此值必須小於或等於序數欄位的值數目上限。

提高資料品質

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。 偏離值可透過修剪 (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、名義或序數欄位的遺漏值。

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，作為次數最少的類別。 請注意，即使原始欄位為字串，新欄位仍會是數值。 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

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作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**」的對應常態分佈值。 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作為母群參數估計值，對應至具有指定之「**最小值**」和「**最大值**」的對應均勻分佈值。 指定「**最大值**」大於「**最小值**」的數目。

連續目標。 這會使用博克斯-考克斯 (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。

附註：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。 為了解譯及使用結果，您必須將預測值轉換回原始比例尺。 如需相關資訊，請參閱 。 如需相關資訊，請參閱 第 14 頁的『反向轉換分數』 。

轉換欄位

若要提升資料的預測能力，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其他所有「轉換欄位」控制項，同時維持選擇。

分類式輸入欄位 您可以使用的選項如下：

- **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，建立較精簡的模型。 相同的類別是根據輸入和目標之間的關係加以識別。 沒有顯著差異 (即 *p* 值大於指定值) 的類別都會被合併。 請指定一個大於 0 且小於或等於 1 的值。 如果所有種類都合併為一個，則會將該欄位的原始版本和衍生版本排除在進一步分析之外，因為它們沒有值作為預測值。
- **若無目標，則根據個數合併稀疏類別**。 如果資料集沒有目標，則您可以選擇合併次序與名義欄位的稀疏類別。 等頻法用於合併含有少於記錄總數之指定最小百分比的類別。 請指定一個大於 0 且小於或等於 100 的值。 預設值為 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。 如果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* 值作為關鍵值之 alpha 的 Scheffe 方法所識別，以判斷同質子集。 請指定一個大於 0 且小於或等於 1 的值。 預設值為 0.05。 如果 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，因為它們沒有值作為預測值。

附註：ADP 中的 binning 和最適 binning 不同。 最佳 binning 使用加密資訊將連續欄位轉換為類別欄位；這需要將資料排序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，這表示 ADP binning 不需要將資料排序，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。

選取和構建

若要提升資料的預測能力，您可以根據現有的欄位來建構新欄位。

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，以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

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，以更符合您的需要與資料。

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，指定要套用至透過「選取」和「建構」設定建構之任何特性的字首名稱。 透過將數值字尾附加到此字首根名稱來建立新名稱。 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 個建構的特性將命名為：feature001 至 feature999，依此類推。

這會確保無論有多少個特性，建構的特性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。

從日期與時間擷取的循環元素。 指定副檔名以套用至從日期與時間擷取的循環元素。

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」對話框而定，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

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作為新欄位。「**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」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 (例如 ID、地址或名稱) 的欄位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

「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」對話框中，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」對話框沒有此控制項，因為若您按一下「**貼上**」，其會將轉換貼到語法視窗作為指令語法。

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」對話框沒有此控制項，因為若您在對話框上方的工具列中按一下「**儲存 XML**」，其會將轉換儲存為 XML。

分析標籤

附註：「互動式資料準備」對話框中的「分析」標籤可讓您檢視建議的轉換。 「自動資料準備」對話框不包含此步驟。

1. 當 ADP 設定 (包括對「目標」、「欄位」及「設定」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」標籤中顯示結果。

「分析」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。

「分析」標籤由兩個畫面組成，主要視圖位於左側，鏈結或輔助視圖位於右側。 主要視圖有三種：

- 欄位處理摘要（預設值）。 如需相關資訊，請參閱主題第 10 頁的『欄位處理摘要』。
- 欄位。 如需相關資訊，請參閱主題第 10 頁的『欄位』。
- 動作摘要。 如需相關資訊，請參閱主題第 11 頁的『動作摘要』。

有四個鏈結/輔助視圖：

- 預測能力 (預設值)。 如需相關資訊，請參閱主題第 11 頁的『預測能力』。
- 欄位表格。 如需相關資訊，請參閱主題第 11 頁的『欄位表格』。
- 欄位詳細資料。 如需相關資訊，請參閱主題第 12 頁的『欄位詳細資料』。
- 動作詳細資料。 如需相關資訊，請參閱主題第 12 頁的『動作詳細資料』。

視圖之間的鏈結

在主要視圖中，表格內加底線的文字會控制鏈結視圖中的顯示。 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。 您最後選取的鏈結會以較暗的顏色顯示，這可協助您識別兩個檢視畫面內容之間的關係。

重設檢視

若要重新顯示原始的「分析」建議並捨棄您對「分析」視圖所做的任何變更，請按一下主視圖畫面下方的「**重設**」。

欄位處理摘要

「欄位處理摘要」表格提供處理的預計整體影響 Snapshot，其中包括特性狀態的變更以及建構的特性數目變更。

請注意，實際上不會建置任何模型，因此在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，您可以顯示個別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。
- 不建議以任何格式使用輸入預測值的數目，無論是原始格式，作為衍生的欄位還是作為建構預測值的輸入。

在加底線的任一「**欄位**」資訊按一下，即可在鏈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」鏈結檢視中。 請參閱第 11 頁的『欄位表格』主題，以取得更多資訊。 **建議在分析中使用的特性**顯示在「預測能力」鏈結視圖中。 請參閱第 11 頁的『預測能力』主題，以取得更多資訊。

欄位

「欄位」主要視圖顯示處理的欄位，以及 ADP 是否建議將它們用於下游模型中。 您可以置換任何欄位的建議；例如，排除結構特性或包含 ADP 建議排除的特性。 如果欄位已轉換，您可以決定是接受建議的轉換還是使用原始版本。

「欄位」視圖包含兩個表格，一個代表目標，一個代表已處理或已建立的預測值。

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；一律使用原始名稱，即使欄位已轉換。
- **測量層級**。 這會顯示代表測量層次的圖示；將滑鼠移至圖示上方可顯示說明資料的標籤（連續、次序、名義等）。

  若目標經過轉換，則**測量層級**欄會反映最終的轉換版本。 附註：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。 依預設，列是以預測能力的遞減順序排序。

對於一般的功能，原始名稱永遠會作為名義稱。 原始和衍生版本的日期/時間欄位會顯示於表格中（以個別列顯示）；表格也會包括建構的預測值。

請注意，表格中顯示的已轉換欄位版本一律代表最終版本。

依預設，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**」方框；接著就會在表格下方顯示這些欄位。

表格包含下列直欄：

- **要使用的版本**。 這會顯示下拉清單，可控制欄位是否用於下游，以及是否使用建議的轉換。 依預設，下拉清單會反映建議。

  對於已轉換的一般預測值，下拉清單有三個選項：**轉換**、**原始**，以及**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，下拉清單是停用的，並且設為「**不使用**」。

  附註：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**測量層級**和**預測能力**設定。
- **名稱**。 每個欄位名稱都是一個鏈結。 在名稱上按一下可以在鏈結的視圖中顯示欄位的相關資訊。 如需相關資訊，請參閱主題第 12 頁的『欄位詳細資料』。
- **測量層級**。 這會顯示代表資料類型的圖示；將滑鼠移至圖示上方可顯示說明資料的標籤（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。 通常，預測能力對於在 ADP 分析內比較預測值非常有用，但是不應在分析中比較預測能力值。

動作摘要

系統會針對自動資料準備所採取的每一個動作，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。 然後，系統會建議將留到最後一個步驟的欄位用於建模，並且過濾出轉換和建構預設值的輸入。

「動作摘要」是一個簡式表格，其中列出 ADP 採取的處理動作。 按一下其中任何加底線的**動作**，便會在鏈結的檢視中顯示更多關於執行動作的詳細資料。 如需相關資訊，請參閱主題第 12 頁的『動作詳細資料』。

附註：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

在第一次執行分析，或是選取「欄位處理摘要」主要視圖的**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」標籤的「欄位名稱」面板選擇的字尾；例如：*_transformed*。

測量層級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

當您在「欄位處理摘要」主視圖中按一下「**目標**」、「**預測值**」或「**未使用的預測值**」時，就會顯示「欄位表格」視圖，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，即使目標已轉換。

  對於轉換版本的一般預測值，名稱反映您在「設定」標籤的「欄位名稱」畫面選擇的字尾；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*。
- **測量層級**。 這會顯示代表資料類型的圖示。

  對於目標，**測量層級**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

當您在「欄位」主視圖中按一下任何「**名稱**」時，就會顯示「欄位詳細資料」。「欄位詳細資料」視圖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歷程和轉換欄位的名稱（如果適用）。

對於每個圖表集合，會以並排的方式顯示兩個版本，以比較套用和未套用轉換的欄位；如果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，只會顯示新預測值的圖表。

附註：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正常曲線，而垂直參照線代表平均值；類別欄位顯示為長條圖。

直方圖標示為顯示標準差及偏斜度；然而，如果值的數目是 2 或更少，或者原始欄位的變異數少於 10-20，則不會顯示偏斜度。

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

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；圖表標籤會顯示百分比。

如果 ADP 執行遺漏值處理，則轉換後的圓餅圖也包括取代值以作為標籤，即使用該值取代遺漏值。

將滑鼠移到圖表上方，以顯示遺漏值計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。 如果目標已轉換，則計算的預測能力與轉換後的目標有關。

附註：若未定義目標或是在主要視圖畫面中按一下目標，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；然而，針對某些動作的特定欄位，可能會執行多個動作。

附註：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」。 如需相關資訊，請參閱主題第 12 頁的『動作詳細資料』。
- **詳細資料**。 所執行處理的清單。 例如，轉換為標準單位。
- **函數**。 此項目僅針對建構的預測值顯示，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

當您在「動作摘要」主要視圖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」鏈結的視圖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，會在鏈結視圖上方使用說明作為標題。 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。

附註：若關閉某個動作或未指定任何目標，則在「動作摘要」主要視圖中按一下該動作時，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。

檢查測量層級表格

此表格會顯示重新分配的欄位數目，內容分為：

- 序數欄位 (排序集合) 重新分配為連續欄位。
- 連續欄位重新分配為序數欄位。
- 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」標籤的「準備輸入與目標」面板中的設定而定，可能是已發現並修整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：

- 博克斯-考克斯 (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）和 Lambda 值的直欄。
- 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

若目標已被 ADP 轉換，則後續的模型會使用已轉換的目標對已轉換的單位進行評分。 為了解譯及使用結果，您必須將預測值轉換回原始比例尺。

1. 若要反向轉換分數，在功能表中選擇：

   轉換 > 準備建模的資料 > 反向轉換分數...
2. 選取欄位以執行反向轉換。 此欄位應包含轉換目標的模型預測值。
3. 指定新欄位的字尾。 這個新欄位將包含未轉換目標的原始比例尺中的模型預測值。
4. 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「自動化資料準備」對話框中儲存的檔案。 如需相關資訊，請參閱主題第 9 頁的『套用並儲存轉換』。

識別異常觀察值

「異常偵測」程序會搜尋以其集群群組標準的差異為基礎的異常個案。 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種類變數上。 每一列都代表一個不同的觀察，且每一行都代表對等群組所依據的不同變數。 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案。 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，用於建立模型的遺漏值處理也許會在計分前套用至檢定資料檔案。

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任何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) 分佈，且每個種類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佈假設偏差的影響，但是要注意這些假設符合的程度。

識別異常觀察值

1. 從功能表中選擇：

   資料 > 識別異常的觀察值...
2. 選取至少一個分析變數。
3. 您也可以選擇一個觀察值 ID 變數，用於標籤輸出。

具有未知測量層級的欄位

若在資料集中出現一或多個未知的變數 (欄位) 測量層級，就會顯示「測量層級」警示。 由於測量層級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層級。

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層級的任何欄位指派預設的測量層級。 若為大型資料集，則讀取時可能需要一些時間。

手動指派。 開啟對話框，以列出具有未知測量層級所有欄位。 您可以使用此對話框以指派測量層級給這些欄位。 您可以在「資料編輯器」的「變數視圖」中指派測量層級。

由於測量層級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層級之前，無法存取對話框來執行此程序。

識別異常的觀察值輸出

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編號、原應變數、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」標籤指定了觀察值 ID 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佈摘要。

- **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性統計量。
- **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性統計量。 如果「選項」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

儲存變數。 這個組別內的控制可讓您將模型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指數值。

- **對等群組**。 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、其變數槓桿值、其本身數值及基準數值。 組合的數量視「選項」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。

匯出模型檔案。 可讓您以 XML 格式儲存模型。

識別異常觀察值的遺漏值：

「遺漏值」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

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。

附註：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群組。

最大原因數。 一個原因會包括變數槓桿值、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》。

最佳 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（自此後稱為「**Binning 輸入變數**」）。 對「supervise（監督）」binning 處理的種類引導變數而言，Bin 資訊都是最適值。 然後系統會使用 Bin 做進一步分析，而不是使用原始的資料值。

範例。 減少變數所帶的不同值個數有若干種用法，其中包括：

- 其他程序的資料需求。 離散化變數可被視為類別變數，以供在需要類別變數的程序中使用。 例如，「交叉表」程序要求所有變數都是種類變數。

- 資料隱密性。 報表經過 bin 處理值而非實際值有助於保護資料來源的隱密性。 「最適 Binning」程序可以引導選擇 Bin。
- 速度效能。 當使用個數減少的不同值時，部分程序較有效率。 例如，使用離散化變數時，可以提升「多項式邏輯斯迴歸 (Multinomial Logistic Regression)」的速度。
- 揭露完全或幾乎完全的資料分隔。

最適 Binning 對視覺化 Binning。 「視覺化 Binning」對話框提供數個自動方法，不使用引導變數而直接建立 Bin。 這些「非監督式」規則對於產生敘述性統計量 (如頻率表格) 非常有用，但是如果您的最終目標是產生預測性模型，則「最適 Binning」較為優越。

輸出。 此程序會為 Bin 產生切點表格，並為每一個 Binning 輸入變數產生敘述性統計量。 此外，您可以將新變數儲存至含有 Binning 輸入變數之經過 bin 處理值的作用中資料集，並儲存 bin 處理規則作為離散化新資料時使用的指令語法。

最適 Binning 資料考量

資料。 此程序預期 Binning 輸入變數是尺度變數、數值變數。 引導變數應為類別引導變數，並且可以為字串或數值。

獲取最佳 Binning

1. 在功能表上，選擇:

   轉換 > 最適 Binning...
2. 選取一個或多個 Binning 輸入變數。
3. 選取引導變數。

依預設不會產生含有經過 bin 處理資料值的變數。 使用儲存標籤來儲存這些變數。

最適 Binning 輸出

「輸出」標籤控制結果的顯示。

- **Bin 的端點**。 顯示每個 Binning 輸入變數的端點集。
- **已經過 bin 處理之變數的敘述性統計量**。 對於每一個 Binning 輸入變數，這個選項會顯示具有有效值的觀察值個數、具有遺漏值的觀察值個數、不同有效值的個數，以及最小值和最大值。 對於引導變數，這個選項會顯示每一個相關 Binning 輸入變數的類別分配。
- **已經過 bin 處理之變數的模型熵**。 對於每一個 Binning 輸入變數，這個選項會顯示有關引導變數的變數之預測準確性測量。

最適 Binning 儲存

儲存變數至作用中資料集。 在進一步分析中，可以使用含有經過 bin 處理資料值的變數代替原始變數。

將 Binning 規則儲存為語法。 產生可用來 bin 處理其他資料集的指令語法。 紀錄規則是根據 Binning 演算法所決定的切點。

最適 Binning 遺漏值

「遺漏值」標籤指定是使用整批刪除法或成對刪除法來處理遺漏值。 一律將使用者遺漏值視為無效。 將原始變數值紀錄到新變數時，使用者遺漏值會轉換成系統遺漏值。

- **成對**。 這個選項運作於每一個引導和 Binning 輸入變數對組。 此程序將利用引導和 Binning 輸入變數上的所有具有非遺漏值的觀察值。
- **整批** 這個選項運作於「變數」標籤上所指定的全部變數。 如果有遺漏觀察值的任何變數，則會排除整個觀察值。

最適 Binning 選項

前置處理。 具有許多不同值的「預先 Binning」Binning 輸入變數可以改善處理時間，而不會太過犧牲最終 Bin 的品質。 最大 Bin 數目對於所建立的 Bin 數目賦予上限。 因此，如果您指定 1000 為最大數目，但是

Binning 輸入變數有少於 1000 個的不同值，則為 Binning 輸入變數所建立的已前置處理 Bin 數目，將等於 Binning 輸入變數中的不同值數目。

稀疏移入的 Bin。 此程序有時可能會產生觀察值極少的 Bin。 下列策略會刪除這些虛擬切點：

對於給定的變數，假設演算法找到 n_{final} 個切點，因此就是 $n_{final}+1$ 個 Bin。 對於 Bin $i = 2, ..., n_{final}$ (第二個最低值 Bin 到第二個最高值 Bin)，計算

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

其中 *sizeof(b)* 是 Bin 中的觀察值個數。

當這個值小於指定的合併臨界值時，無論哪個具有較低的類別資訊熵，b_i 會被視為稀疏移入，並與 b_{i-1} 或 b_{i+1} 合併。

此程序會進行單一透通 Bin。

Bin 端點。 這個選項指定間隔的下限是如何定義。 由於此程序會自動決定切點的值，因此這主要是取決於喜好設定。

第一個 (最低) / 最後一個 (最高) Bin。 這些選項指定每一個 Binning 輸入變數的最小和最大切點是如何定義。 一般而言，此程序是假設 Binning 輸入變數可以取實數行上的任何值，但是如果您有理論上或實際上的原因得限制範圍，可以用最低值 / 最高值來約束它。

OPTIMAL BINNING 指令的其他功能

指令語法語言也可以讓您：

- 透過相等頻率方法執行非監督式 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》。

注意事項

本資訊係針對 IBM 在美國所提供之產品與服務所開發。 IBM 可能會提供此資料的其他語言版本。 然而， 貴客戶可能需要擁有該語言的產品或產品版本副本，才能進行存取。

IBM 可能不會在其他國家或地區提供本文件所討論的產品、服務或特性。 請洽詢當地的 IBM 業務代表，以取得當地目前提供的產品和服務之相關資訊。 本文件在提及 IBM 的產品、程式或服務時，不表示或暗示只能使用 IBM 的產品、程式或服務。 只要未侵犯 IBM 之智慧財產權，任何功能相當之產品、程式或服務皆可取代 IBM 之產品、程式或服務。 不過，任何非 IBM 之產品、程式或 服務，使用者必須自行負責作業之評估和驗證責任。

本文件所說明之主題內容，IBM 可能擁有其專利或專利申請案。 提供本文件不代表授與這些專利的授權。 您可以用書面方式來查詢授權，來函請寄到：

IBM Director of Licensing

IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

若要查詢有關雙位元組 (DBCS) 資訊的授權事宜，請洽詢所在國家或地區的 IBM 智慧財產部門，或書面提出授權查詢，來函請寄到：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

International Business Machines Corporation 只依「現況」提供本出版品，不提供任何明示或默示之保證， 其中包括且不限於不侵權、可商用性或特定目的之適用性的隱含保證。 有些地區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。

本參考資訊中可能會有技術上或排版印刷上的訛誤。 因此，IBM 會定期修訂；並將修訂後的內容納入新版中。 IBM 隨時會改進及/或變更本出版品所提及的產品及/或程式，不另行通知。

本資訊中任何對非 IBM 網站的敘述僅供參考，IBM 對該網站並不提供任何保證。 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。

如果本程式之獲授權人為了 (i) 在個別建立的程式和其他程式（包括本程式）之間交換資訊，以及 (ii) 相互使用所交換的資訊，因而需要相關的資訊，請洽詢：

IBM Director of Licensing

IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

這些資訊可能可以使用，但必須遵循適當的條款，在某些情況中需要付費。

IBM 基於雙方之「IBM 客戶合約」、「IBM 國際程式授權合約（或任何同等合約）條款，提供本文件所提及的授權程式與其所有適用的授權資料。

本文件中引用的效能資料及用戶範例僅供敘述之目的。 實際效能結果可能會依據特定配置和作業條件而有所不同。

本文件所提及之非 IBM 產品資訊，係取自產品供應商，或其發佈的聲明或其他公開管道。 IBM 並未測試過這些產品，也無法確認這些非 IBM 產品的執行效能、 相容性或任何對產品的其他主張是否完全無誤。 有關非 IBM 產品功能之問題，應直接洽詢產品供應商。

所有關於 IBM 未來方針或目的之聲明，隨時可能更改或撤銷，不必另行通知，且僅代表目標與主旨。

本資訊含有日常企業運作所用之資料和報告範例。 為了盡可能詳盡說明，這些範例都包括個人、公司、品牌及產品的名稱。 所有這些名稱全為虛構，任何與實際人員或商業企業類似之處，純屬巧合。

著作權授權:

本資訊含有原始語言之範例應用程式，用以說明各作業平台中的程式設計技術。 貴客戶可以為了研發、使用、銷售或散布符合範例應用程式所適用的作業平台之應用程式介面的應用程式，以任何形式複製、修改及散布這些範例程式，不必向 IBM 付費。 這些範例並未在所有情況下完整測試。 因此，IBM 不保證或暗示這些程式的可靠性、服務性或功能。 這些程式範例以「現狀」提供，且無任何保證。 IBM 對因使用這些程式範例而產生的任何損害概不負責。

這些範例程式或任何衍產生果的每份複本或任何部分，都必須依照下列方式併入著作權聲明:

商標

IBM、IBM 標誌及 ibm.com 是 International Business Machines Corp. 在世界許多管轄區註冊的商標或註冊商標。 其他產品及服務名稱可能是 IBM 或其他公司的商標。 IBM 商標的最新清單可在 Web 的 "Copyright and trademark information" 中找到，網址為 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 及 Pentium 是 Intel Corporation 在美國及（或）其他國家或地區商標或註冊商標。

Linux 是 Linus Torvalds 在美國及/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。

Java 及所有 Java 型商標及標誌是 Oracle 及/或附屬公司的商標或註冊商標。

索引

Special Characters

B

M