IBM SPSS Data Preparation 26

附註

使用此資訊和支援的產品之前，請先閱讀第 5 頁的『注意事項』中的資訊。

產品資訊

若新版本未聲明，則此版本便適用於 IBM® SPSS Statistics 26.0.0 版以及之後發行的所有版本和修正。

目錄

資料準備

下列資料準備功能包含在 SPSS® Statistics Professional Edition 或「資料準備」選項中。

資料準備簡介

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

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

使用資料準備程序

是否使用資料準備程序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。

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

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

Statistics

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指數，及當觀察值被視為異常時影響最大之變數的變數影響數值。

資料考量

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

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

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

假設。 演算法假設所有變數都是非常數且獨立，並假設所有觀察值在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，且每個類別變數都假設具有多項式分配。經驗內部檢定指出此程序很少受到獨立性假設及分配假設偏差的影響，但是要注意這些假設符合的程度。

識別異常觀察值

1. 在功能表上，選擇：

   資料 > **識別異常的觀察值...**
2. 至少要選取一個分析變數。
3. 您也可以選擇一個觀察值 ID 變數，用於標記輸出。
4. 按一下**套用**。

測試層級不明的欄位

如果資料集中一或多個變數（欄位）的測量層級不明，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

列出其測量層級不明的所有欄位。您可以將測量層級指派給那些欄位。您也可以在「資料編輯器」的「變數清單」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，無法執行此程序。

識別異常觀察值：輸出

「輸出」對話框提供選項用來產生表狀輸出。

異常觀察值及它們為什麼被視為異常的原因清單

選取的話，使用此選項會產生三個表格：

- 異常觀察值指數會列出被識別為異常的觀察值，並顯示它們的對應異常指數值。
- 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、原因變數、變數影響數值、變數數值及變數的基準。

所有的表格皆以遞減的順序由異常指數排列。此外，如果在「變數」對話框上指定了觀察值 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配摘要。

對等組別基準

這個選項顯示連續變數基準表格（如果分析中使用任何連續變數）及類別變數基準表格（如果分析中使用任何類別變數）。連續變數基準表格顯示每個對等組別中各連續變數的平均數及基準差。類別變數基準表格顯示每個對等組別中各類別變數的眾數 (最普遍的類別)、次數及次數百分比。分析時會將連續變數的平均數及類別變數的眾數當成標基準值使用。

異常指標

異常指數摘要會顯示被視為異常程度最高之觀察值的異常指數描述性統計量。

依分析變數排列的發生原因

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描述性統計量。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。

已處理的觀察值

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。

識別異常觀察值：儲存

「儲存」對話框提供變數與模型儲存選項。

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。

異常指數

以指定的變數名稱儲存每個觀察值的異常指數值。

對等組別

以指定的變數根名稱儲存每個觀察值的對等組別 ID、觀察值個數及大小百分比。例如，如果已經指定根名稱「*Peer*」，則會產生「*Peerid*」、「*PeerSize*」，及「*PeerPctSize*」等變數。「*Peerid*」是觀察值的對等組別 ID，「*PeerSize*」是組別的大小，「*PeerPctSize*」是組別大小的百分比。

原因　以指定的根名稱儲存推理變數的組合。推理變數組合包括作為原因的變數名稱、其變數槓桿值、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。

替換名稱或根名稱相同的現有變數

選取的話，會替換其名稱與要儲存的變數相衝突的現有變數。

匯出模型檔

可讓您以外部 XML 檔形式儲存模型。

識別異常觀察值：遺漏值

「遺漏值」對話框用於控制使用者遺漏及系統遺漏值的處理。

從分析中排除遺漏值

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。

識別異常觀察值：選項

「選項」對話框包括的設定可用於異常觀察值準則以及用於定義對等組別個數的範圍。

用於識別異常觀察值的準則

下列設定決定異常清單將包括多少觀察值。

異常指數值最高的觀察值百分比

請指定一個小於或等於 100 的正數。

異常指數值最高的固定數量觀察值

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。

只識別其異常指數值符合或超過最低值的觀察值

指定一個非負數的數字。如果觀察值的異常指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指數值都大於或等於 2。

對等組別的個數

這個程序會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。

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組別。

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 指令其他特性

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》。

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 可能會以其他語言提供本資料。但是，您可能需要具有該語言的產品或產品版本，才能存取該產品。

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, MD-NC119
Armonk, NY 10504-1785
US

對於與雙位元組 (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.
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。

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。

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。

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 Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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。

本文件中引用的效能資料及用戶範例僅供敘述之目的。特定配置及作業條件下的實際效能結果可能不同。

本文件所提及之非 IBM 產品資訊，取自產品的供應商，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 產品的功能問題應直接洽詢該產品供應商。

關於 IBM 未來方針或意圖的所有聲明僅代表目標或目的，得依規定未另行通知即變更或撤銷。

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。這些名稱全為虛構，如與實際人員或企業之名稱有所雷同，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之程式設計技術。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。故 IBM 不保證或默示保證這些樣本程式之可靠性、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生成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 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。

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。

Java 和所有以 Java 為基礎的商標及標誌是 Oracle 及（或）其子公司的商標或註冊商標。

索引

索引順序以中文字，英文字，及特
殊符號之次序排列。

〔十劃〕

〔十一劃〕

〔十四劃〕

〔十六劃〕

〔十九劃〕

Printed in Taiwan