當代教育研究季刊　第二十三卷　第四期，2015 年 12 月，頁 121-153
DOI : 10.6151/CERQ.2015.2304.05


# 運用開放軟體 R 處理大型教育資料庫

紀馥安 [1*]　許清芳 [2]

## 摘　要

目前教育工作者處理大型教育資料庫常用的工具多為商用的統計軟體。資料管理與分析的流程因此不易與其他研究者分享，結果也較難被覆驗。信度不能累積的研究成效也不容易幫助指引教育政策的改進方向。

本研究的主旨在示範如何運用開放軟體 R 管理與分析大型教育資料庫的檔案。案例為，包括本國在內共十五個國家，十五歲學生數學成就表現與多個影響該表現的相關因素（PISA 2012）。範例分析，除了使用 R 的基本功能外，特別嵌入 Caro 與 Biecek（2014）針對處理大型教育資料庫發展的 intsvy 套件並結合 maptools 等相關套件將分析結果呈現在地圖上以利比較。文中提及的檔案管理及資料分析過程皆可使用提供在附錄的 R 程式碼複製。

**關鍵詞**： intsvy、PISA、R、國際大型教育測驗資料

* 紀馥安（通訊作者），國立成功大學教育研究所博士生
電子郵件：u38011033@mail.ncku.edu.tw
許清芳，國立成功大學教育研究所教授
電子郵件：csheu@mail.ncku.edu.tw
投稿日期：2015 年 04 月 30 日；修正日期：2015 年 07 月 20 日；接受日期：2015 年 09 月 16 日

December, 2015, Vol. 23, No. 4, pp. 121-153
DOI : 10.6151/CERQ.2015.2304.05

# USING R TO ANALYZE INTERNATIONAL LARGE-SCALE EDUCATIONAL ASSESSMENT DATA

**Fu-An Chi [1*] Ching-Fan Sheu [2]**

## ABSTRACT

Currently, researchers use mostly commercial software to analyze educational data. Researchers who do not have access to these commercial software can not easily replicate results others have reported in publications. Consequently, computational findings often do not cumulate reliably to offer valid guidelines for educational policy making.

This paper demonstrates how to use the free, open-source R computing environment (R Core Team, 2015) to manage and analyze international large-scale educational assessment data. The example is comprised of mathematics achievement and covariates for 15-year-old students from 15 countries, including Taiwan (PISA 2012). In particular, the "intsvy" package developed by Caro and Biecek (2014) is applied to manage the data and the "maptools" package (Bivand & Lewin-Koh, 2015) is used to link numerical summaries to geographical boundaries of countries examined in the illustration. R code to perform analysis of data from PISA detailed in the paper is provided to facilitate reproducibility.

*Keywords:* *intsvy, PISA, R, international large-scale educational assessment data*

## 壹、前言

隨著近年來世界各國積極參與國際大型學生學習成就評比，像是國際教育成就評鑑協會（International Association for the Evaluation of

* Fu-An Chi (corresponding author), Doctoral Student, Institute of Education, National Cheng Kung University.
E-mail: u38011033@mail.ncku.edu.tw
Ching-Fan Sheu, Professor, Institute of Education, National Cheng Kung University.
E-mail: csheu@mail.ncku.edu.tw
Manuscript received: April 30, 2015; Modified: July 20, 2015; Accepted: September 16, 2015

Educational Achievement，簡稱 IEA）所舉辦的國際數學與科學教育成就趨勢調查（Trends in International Mathematics and Science Study，簡稱 TIMSS），自 1995 年起每四年一次針對四年級與八年級學生的數學與科學成就進行調查，藉此瞭解各國教育制度、學校課程、教師素質與家庭環境等相關因素（Martin, Mullis, & Foy, 2008）。同樣 IEA 所舉辦的促進國際閱讀素養研究（Progress in International Reading Literacy Study，簡稱 PIRLS），每五年一次針對四年級學生的閱讀素養進行評量，根據閱讀目的、理解過程、閱讀行為與學習態度來得知各國學生閱讀素養的程度（柯華葳、詹益綾、丘嘉慧，2013）。另外經濟合作暨發展組織（Organisation for Economic Co-operation and Development，簡稱 OECD）自 1997 年起舉行的國際學生能力評量計劃（the Programme for International Student Assessment，簡稱 PISA）（臺灣 PISA 國家研究中心，2015），主要針對 15 歲學生閱讀、數學與科學進行評量（臺灣 PISA 國家研究中心，2014），透過學校課程，學校氣氛、社團活動、家庭環境等方面（臺灣 PISA 國家研究中心，2015），看見學生在學業學習與問題解決上的素養（臺灣 PISA 國家研究中心，2014）。以及 OECD 整合 IALS 與 ALL 兩次大型國際成人素養調查結果的國際成人能力評量計畫（Programme for the International Assessment of Adult Competencies，簡稱 PIAAC），此計畫於 2011 年開始進行，主要針對教育投資轉換為成人關鍵能力與社會經濟成效的測量（Organisation for Economic Co-operation and Development [OECD], 2010; Schleicher, 2008），這些國際評比促使大型教育資料庫的相關研究日益漸增，例如：Rutkowski、Gonzalez、Joncas 與 von Davier（2010）基於 2009 年 8 月 Wilson Education Full Text 資料庫針對 PISA、TIMSS 與 PIRLS 的研究指出，從 1995 年到 1999 年有 340 篇已發表的文章、2000 年到 2004 年間有 556 篇、從 2005 年到 2009 年更有 851 篇。因此可以推斷過去十年在學術界越來越多研究者關注與應用國際大型測驗的數據（Liou & Hung, 2014）。

在臺灣教育研究領域中，研究者使用 SPSS 進行統計分析的使用率高達 75.5%，其次是 SAS（麥馨月，2004）。而這些軟體雖然在資料處理、介面操作、圖表繪製上各有優點，但這些軟體基本上都需要付費才能取

得，對於研究者而言無疑是一種經濟上的負擔，以及大型教育資料庫所提供的數據多設定為這兩種類型的檔案，無形中侷限了研究者對大型教育資料庫的分析，再加上研究者從大型教育資料庫下載原始數據後經常花費不少時間在整理數據上，例如：需要先在 SPSS 上挑選變項與國家、處理反向題等問題。此外複現是對科學主張作評判的基本標準，這種再現性的標準要求數據與用來分析數據的電腦程式碼提供給其他人（Peng, 2011）。但目前教育工作者處理大型教育資料庫常用的工具多為商用的統計軟體，資料管理與分析的流程因此不易與其他研究者分享，結果也較難被覆驗，像這樣信度不能累積的研究成效也不容易幫助指引教育政策的改進方向。

因此由 Ross Ihaka 與 Robert Gentleman 所開發的 R（Ihaka & Gentleman, 1996）在學術界有日漸受到重視的趨勢，因為 R 具備免費、容易取得共享資源、支援 Windows、LINUX…等多種作業系統、擁有強大的繪圖功能，包含了 2D 與 3D 的繪圖、內建許多簡單易學的統計語法、易於使用者擴展所編寫的函數、豐富且活躍更新擴充套件（Packages）、對各種資料格式的輸入與輸出提供支援、並與其他程式語言（例如：C、C++、JAVA、Fortran、Perl 與 Python）有高度的相容性、對常用軟體 SAS 與 SPSS 可經由安裝 Plug-in 程式直接讀取 R 的語法等優點，可幫助研究者解決過去使用其他軟體時所遭遇的不便。2014 年 Daniel Caro 與 Przemyslaw Biecek 開發出 R 的 intsvy 套件，即可涵蓋 TIMSS、PIRLS、PISA 與 PIAAC 四個大型教育資料庫，它不僅有合併數據、命名變項與國家的名稱以及直接將所選擇學生、家長、學校、教師教學的變項與所選擇的國家直接導入到 R 等整理數據的功能，同時也具有平均數、標準差、迴歸估計、相關係數、次數表、計算點估計與標準誤的複雜樣本設計（即重複抽樣加權值）以及旋轉測試形式（Rotated test forms，即成績似真值）等分析數據的功能（Caro & Biecek, 2014）。使用 intsvy 套件不僅可幫助研究者專心在 R 裡面整理數據，同時分析多國數據，甚至是繪製出多國的圖表，藉此減少研究者許多不必要的人力與時間。intsvy 套件也因此逐漸受到教育研究者的關注與應用，像是 Benton（2014）就使用 intsvy 套件針對 PISA 2009 的資料進行跨國比較研究。

本研究的主旨在示範如何運用開放軟體 R 管理與分析大型教育資料庫的檔案。案例為，包括本國在內共十五個國家，十五歲學生數學成就表現與多個影響該表現的相關因素（PISA 2012）。範例分析，除了使用 R 的基本功能外，特別嵌入 Caro 與 Biecek（2014）針對處理大型教育資料庫發展的 intsvy 套件並結合 maptools 等相關套件將分析結果呈現在地圖上以利比較。文中提及的檔案管理及資料分析過程皆可使用提供在附錄的 R 程式碼複製。

# 貳、文獻探討

## 一、過去處理大型教育資料庫的常用統計分析軟體

80 年代因電腦處理速度提升且功能複雜，使套裝軟體開始盛行，進而替代程式語言，成為應用於社會與行為科學研究的軟體，其中 SPSS 最受到社會科學研究者廣泛使用。教育研究領域中，麥馨月（2004）針對民國 45 年至 92 年間臺灣師範大學、政治大學與高雄師範大學教育研究所所通過的 2082 篇學位論文進行分析的「臺灣教育研究的發展與趨勢－以教育研究所學位論文為研究對象」這篇研究中指出，長久以來佔了 75.5% 使用率的 SPSS 一直被大部分研究者用來進行統計分析，其後依序為 SAS 及 BMDP，但近年來新型統計軟體的出現，使得「其他電腦統計軟體」的運用比例也逐漸增加。這三間學校在選擇電腦統計軟體的排序接近一致，大多使用 SPSS 來進行研究上的統計分析。另外博士、碩士論文在使用電腦統計軟體的排序也是趨近相同，多採用 SPSS 進行統計分析，其後依序 SAS、其他電腦統計軟體、BMDP。學位論文在進行統計分析時，大部分只採單一種電腦統計軟體，多數皆選擇 SPSS 作為統計軟體工具。由此可知，SPSS 在教育研究領域中是大部分研究者最常使用的電腦統計軟體。

## 二、R 的介紹

R 是由 Ross Ihaka 與 Robert Gentleman 所開發（Ihaka & Gentleman, 1996）。但 R 可追朔至 70 年代貝爾實驗室的一個研究專案，此專案為了發展一套適於統計研究員的互動式統計分析軟體而開發出 S 語言，即 S-

Plus 的基礎。直到 90 年代初期，在奧克蘭大學任教的 Ross Ihaka 與 Robert Gentleman 因處理學校沒有適合的統計軟體可使用於麥金塔電腦上，而仿擬 S 語言與 Scheme 語言的結構，即 S 的語法加上 Scheme 的基本語義，最後發展出可供在統計教學時可使用的 R。但在發展初期只有 Mac 電腦的版本，直到研發 R 的核心團隊於 1997 年中期後成立，才將開放原始碼的 R 移到不同的作業系統，同時也加入自由軟體基金會（Free Software Foundation，簡稱 FSF）的 GNU 計畫。因此讓 R 有其它商業統計軟體所沒有的特性，像是免費、原始碼開放與功能可擴充等，以致 R 逐漸受到學術界的重視。

R 是一套利於統計計算與圖表繪製的免費軟體系統，它提供一個整合良好的數學計算環境，與 MATLAB、Visual Basic 以及 Java Script 同為直譯式語言且操作方式類似，R 因執行速度快、可立即得到輸出結果、語法與 C 語言相似、語義主要是函數設計語言等特質，因此直接輸入函數即可完成數學運算，這樣的好處有益於統計計算與圖表繪製（R Core Team, 2015）。此外 R 對各種資料格式的輸入與輸出提供支援，與其他程式語言（例如：C、C++、JAVA、Fortran、Perl 與 Python）有高度的相容性，對於常用軟體 SPSS 與 SAS 也能藉由安裝 Plug-in 程式來直接讀取 R 的語法。

由此可知，R 具備以下有利於研究者使用的優點：（一）免費且容易取得共享資源。（二）支援 Windows、LINUX…等多種作業系統。（三）擁有強大的繪圖功能，包含了 2D 與 3D 的繪圖。（四）使用者可利用方便且完善的內建協助系統搜尋指令。（五）R 內建許多簡單易學的統計語法，且易於使用者擴展所編寫的函數。（六）豐富且活躍更新擴充套件（Packages）。（七）提供數個社群資源讓使用者討論、發問或交流學習。（八）所有函數、資料、變項、運算與結果皆可以物件形式儲存在電腦，之後再藉由運算或函數對已儲存在電腦中的物件進行操作。

## 三、R 在其它領域與大型資料庫的應用

R（R Core Team, 2015）已被各種領域視為統計分析的用途。因為 R 是一套專業的統計分析與圖表呈現的自由軟體（Chambers, 2009; R Core Team, 2015），具備快速更新、可從 R Project 網站免費下載使用的優點（Fox, 2009）。R 的綜合典藏網（Comprehensive R Archive Network，簡稱 CRAN），

CRAN 不僅附有 R 的執行檔、原始碼與說明檔，也收納各類用戶所撰寫的套件，目前全球已有超過一百個 CRAN 鏡像站，在 CRAN 網站上的主題列表可知 R 在各領域已被廣泛的使用，例如：財務分析、遺傳學、高性能計算、機器學習、醫學影像、社會科學與空間統計等領域，人工智能、地震模擬與動力運算等特殊應用皆有各自對應的擴充套件（The Comprehensive R Archive Network），因此受到廣大使用者的支持。全球 R 的使用者可自行開發套件的擴充（Fox, 2009），以致 R 應用於各種領域的統計運算。例如：系統生物（Gentleman et al., 2004）、化學領域（Versteeg, Richardson, & Rowe, 2006; Vidal, Thormann, & Pons, 2005），此外許多重要且規模龐大的資料也利用 R 來分析，像是奧地利與英國的選舉預測，處理基因表現數據的系統，以及處理大腦影像的時間序列（Ripley, 2001）。

近年來臺灣也開始有運用 R 分析大型資料庫數據的研究，像是陳靜怡（2012）的研究分析主要是以 R 與 Bioconductor 套件作為操作介面，結合 Array Express 資料庫與 KEGG 資料庫，運用篩選差異性表達基因與基因組富集來分析數據。葉信伶、鄒惠貞、江宏哲、江博煌與劉德明（2013）採用衛生署等相關單位所提供的死因資料庫，利用地理群聚分析及集群點熱圖（Cluster Heat Map）的階層式分層法，以 R 呈現及時產生疾病熱點分析功能的風險。然而 R 應用在大型教育資料庫的研究尚未普及化，身為量化研究的研究者應對其它統計軟體多加認識甚至運用，而不是僅限於對一種統計軟體的瞭解，因此以下介紹 R 可處理大型教育資料庫的 intsvy 套件。

## 參、intsvy 套件的介紹

首先說明一般在 R 進行迴歸分析的做法，以參與 PISA 2012 年臺灣（TAP）的 15 歲學生數學成就表現（MATH）以及家中藏書量（ST28Q01）為例，以 read.csv( ) 讀取放置在電腦“文件”中 PISA2012TAP 的 EXCEL 檔案（資料取得可參考下方範例的說明），其中←代表將右邊的資料存入左邊的檔案，header = TRUE 是指顯示變項的名稱，na.string = "."是出現遺漏值的話，以 . 作為代替。str ( ) 用來檢視資料結構，as.numeric( ) 將

家中藏書量的變項型態轉換成數值（numeric），lm( ) 則是迴歸分析的函數，其中 ~ 的左邊放依變項，右邊放自變項，data = 是指取得存有資料的檔案，再以 summary( ) 檢視該迴歸分析的結果，程式碼見附錄一。

intsvy 是由 Daniel Caro 與 Przemyslaw Biecek 針對 TIMSS、PIRLS、PISA 與 PIAAC 四個大型教育資料庫所開發的套件，具有整理數據與分析數據的功能（Caro & Biecek, 2014）。以下本研究使用 PISA 為例，介紹 intsvy 套件的各種函數：

## 一、數據的選擇與合併

pisa.select.merge 這個函數可從 PISA 提供的數據檔中直接選取與合併所需的檔案與變項，所有的成績與加權的部分都已被預設選取。

## 二、次數表的計算與繪製

pisa.table 這個函數可針對類別變項產生包括百分比與標準誤的次數表。plot 這個函數可將 pisa.table 這個函數產生的次數表進行圖形化。需要注意的是 plot 這個函數目前僅限使用在 PISA 與 PIAAC 這兩個資料庫所提供的數據檔。

## 三、變項平均數的計算與繪製

pisa.mean 這個函數可計算觀察變項（不包括似真值）的平均數與標準誤。pisa.mean.pv 這個函數則可使用五個似真值計算平均成績與標準誤。plot 這個函數可將 pisa.mean 與 pisa.mean.pv 這兩個函數產生的平均數進行圖形化，需要注意的是 plot 這個函數目前僅限使用在 PISA 與 PIAAC 這兩個資料庫所提供的數據檔。

## 四、迴歸分析的計算與繪製

pisa.reg.pv 這個函數可執行有似真值與重複抽樣加權值的線性迴歸分析（OLS）。plot 這個函數可將 pisa.reg.pv 這個函數產生的迴歸分析結果進行圖形化，需要注意的是 plot 這個函數目前僅限使用在 PISA 與 PIAAC 這兩個資料庫所提供的數據檔。

# 肆、範例

本研究旨在示範如何運用開放軟體 R 管理與分析大型教育資料庫的檔案，案例中包括參與 PISA 2012 年臺灣（TAP）與其他 14 個國家，包括澳洲（AUS）、加拿大（CAN）、德國（DEU）、西班牙（ESP）、芬蘭（FIN）、法國（FRA）、英國（GBR）、印尼（IDN）、義大利（ITA）、日本（JPN）、墨西哥（MEX）、馬來西亞（MYS）、紐西蘭（NZL）、美國（USA）的 15 歲學生數學成就表現（MATH）以及影響該表現的學生性別（ST04Q01；Gender）、家中藏書量（ST28Q01；Book）、母親教育程度（ST13Q01；Mother）、家中擁有物（ST26Q01-ST26Q14；Possessions）等相關因素，範例資料取自 PISA 2012 的學生問卷（下載資料的網址：http://pisa2012.acer.edu.au/downloads.php），取得 PISA 2012 的讀取學生問卷資料 SPSS 語法（SPSS syntax to read in student questionnaire data file）與學生問卷資料檔（Student questionnaire data file）後，在讀取學生問卷資料 SPSS 語法中的 DATA LIST FILE = "C:\XXX\INT_STU12_DEC03.txt" /，輸入學生問卷資料檔存放在電腦的位置，以及在讀取學生問卷資料 SPSS 語法的最後一行輸入 SAVE OUTFILE = "C:\XXX\INT_STU12_DEC03.sav".，以取得 PISA 2012 的學生問卷資料。範例分析除了使用 R 的基本功能外，特別嵌入 Caro 與 Biecek（2014）針對處理大型教育資料庫發展的 intsvy 套件並結合 maptools 等相關套件將分析結果呈現在地圖上以利比較。文中提及的檔案管理及資料分析過程皆可使用提供在附錄的 R 程式碼複製，有關 R 的操作則參考鄭中平與許清芳（2015）的「R 在行為科學之應用」一書。

首先安裝與載入 intsvy 套件，以 pisa.select.merge( ) 取得 PISA 2012 學生檔案及研究所需的變項與國家，使用 str( ) 檢視資料結構，結果如表 1，程式碼見附錄二。從表 1 可見，所選的國家、變項以及已被預設選取的全部成績與加權皆包含在內。再利用 as.factor( )、as.numeric( ) 將國家（CNT）的變項型態從字串（character）轉換成因子（factor），將學生性別（ST04Q01）、家中藏書量（ST28Q01）、母親教育程度（ST13Q01）、家中擁有物（ST26Q01-ST26Q14）從因子（factor）轉換成數值（numeric）。

表 1

**資料結構**

| | | | | | | |
|---|---|---|---|---|---|---|
| ‘data.frame’ | : | 189804 obs. of 152 variables: | | | | |
| $ CNT | : | chr | “AUS” | “AUS” | “AUS” | “AUS”… |
| $ SCHOOLID | : | chr | “0000001” | “0000001” | “0000001” | “0000001”… |
| $ STIDSTD | : | chr | “00001” | “00002” | “00003” | “00004”… |
| $ ST04Q01 | : | Factor | w/ 2 levels | “Female”, ”Male”: 1 1 1 2 2 1 1 2 2 2 … | | |
| $ ST13Q01 | : | Factor | w/ 5 levels | “<ISCED level 3A>”, . . : 1 3 NA 1 1 2 1 . . . | | |
| $ ST26Q01 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 2 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q02 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q03 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q04 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q05 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 2 1 1 1 1 1 1 1 … | | |
| $ ST26Q06 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q07 | : | Factor | w/ 2 levels | “Yes”, “No”: 2 2 1 1 2 2 1 2 1 1 … | | |
| $ ST26Q08 | : | Factor | w/ 2 levels | “Yes”, “No”: 2 2 1 1 2 2 1 2 1 1 … | | |
| $ ST26Q09 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 2 1 1 2 1 1 1 1 1 … | | |
| $ ST26Q10 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q11 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 2 2 2 1 2 2 2 1 1 … | | |
| $ ST26Q12 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST26Q13 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 2 … | | |
| $ ST26Q14 | : | Factor | w/ 2 levels | “Yes”, “No”: 1 1 1 1 1 1 1 1 1 1 … | | |
| $ ST28Q01 | : | Factor | w/ 6 levels | “0-10 books”, “11-25 books”, .. 5 3 5 5 3 … | | |
| $ PV1MATH | : | num | 562 | 565 | 507 | 602 | 520… |
| $ PV2MATH | : | num | 569 | 557 | 547 | 594 | 507… |
| $ PV3MATH | : | num | 555 | 553 | 511 | 552 | 501… |
| $ PV4MATH | : | num | 579 | 538 | 454 | 526 | 521… |
| $ PV5MATH | : | num | 548 | 573 | 546 | 619 | 547… |

把原本國家名稱 CNT 改成 Country_ID、原本學生性別名稱 ST04Q01 改成 Gender、原本家中藏書量名稱 ST28Q01 改成 Book，原本母親教育程度名稱 ST13Q01 改成 Mother，原本家中擁有物名稱 ST26Q01-ST26Q14 改成 Possessions。其中將學生性別的原本選項從女生是 1、男生是 2，改成女生是 0、男生是 1，再利用 as.factor( ) 把學生性別從數值轉換成因子，設定女生是 F、男生是 M；家中藏書量則依原本設定 0-10 本書是 1、11-25 本書是 2、26-100 本書是 3、101-200 本書是 4、201-500 本書是 5、500 本書以上是 6；母親教育程度原本選項從高中畢業是 1、高職畢業或五專（不含最後 2 年）是 2、國中畢業是 3、國小畢業是 4、國小肄業是 5，改成高中畢業是 5、高職畢業或五專（不含最後 2 年）是 4、國中畢業是 3、國小畢業是 2、國小肄業是 1；家中擁有物原本選項從 Yes 是 1、No 是 2，

改成 Yes 是 1、No 是 0，並將家中擁有物 14 題的選項加總，最後重新輸入 str( ) 檢視資料結構，結果如表 2，程式碼見附錄三。從表 2 可見，轉換後的各變項型態、名稱以及選項皆已符合上述的設定。

為了瞭解臺灣與其他十四個國家男、女學生的家中藏書量次數、百分比等分配情形，使用 pisa.table( ) 呈現該情形的次數表，以 head( ) 檢視前六行結果，結果如表 3。利用 plot( ) 繪製該次數表，其中 na.omit( ) 是處理資料表格中的遺漏值，stacked( ) 是指資料是否要以堆疊的方式呈現，結果如圖 1，程式碼見附錄四。從圖 1 可知，德國（DEU）不論是男（1）、女學生（0）的家中藏書量有 500 本書以上（6）的人數皆多於其它十四個

表 2

**轉換變項型態、選項、名稱後的資料結構**

| | | | | | | |
|---|---|---|---|---|---|---|
| 'data.frame' | : | 189804 obs. of 157 variables: | | | | |
| $ CNT | : | Factor | w/ 15 levels | "AUS" | "CAN" | "DEU",.. : 1 1 1 1 … |
| $ SCHOOLID | : | chr | "0000001" | "0000001" | "0000001" | "0000001"… |
| $ STIDSTD | : | chr | "00001" | "00002" | "00003" | "00004"… |
| $ ST04Q01 | : | num | 0 0 0 1 1 0 0 1 1 1 … | | | |
| $ ST13Q01 | : | num | 5 3 NA 5 5 4 5 5 5 5 … | | | |
| $ ST26Q01 | : | num | 1 0 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q02 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q03 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q04 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q05 | : | num | 1 1 0 1 1 1 1 1 1 1 … | | | |
| $ ST26Q06 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q07 | : | num | 0 0 1 1 0 0 1 0 1 1 … | | | |
| $ ST26Q08 | : | num | 0 0 1 1 0 0 1 0 1 1 … | | | |
| $ ST26Q09 | : | num | 1 0 1 1 0 1 1 1 1 1 … | | | |
| $ ST26Q10 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q11 | : | num | 1 0 0 0 1 0 0 0 1 1 … | | | |
| $ ST26Q12 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST26Q13 | : | num | 1 1 1 1 1 1 1 1 1 0 … | | | |
| $ ST26Q14 | : | num | 1 1 1 1 1 1 1 1 1 1 … | | | |
| $ ST28Q01 | : | num | 5 3 5 5 3 3 6 5 4 3 … | | | |
| $ PV1MATH | : | num | 562 | 565 | 507 | 602 520… |
| $ PV2MATH | : | num | 569 | 557 | 547 | 594 507… |
| $ PV3MATH | : | num | 555 | 553 | 511 | 552 501… |
| $ PV4MATH | : | num | 579 | 538 | 454 | 526 521… |
| $ PV5MATH | : | num | 548 | 573 | 546 | 619 547… |

表 3

**臺灣與其它十四個國家男、女學生家中藏書量的前六筆次數分配**

| | Country_ID | Gender | Book | Freq | Percentage | Std.err. |
|---|---|---|---|---|---|---|
| 1 | AUS | F | 1 | 679 | 8.93 | 0.48 |
| 2 | AUS | F | 2 | 898 | 12.42 | 0.50 |
| 3 | AUS | F | 3 | 2039 | 29.95 | 0.60 |
| 4 | AUS | F | 4 | 1431 | 21.30 | 0.59 |
| 5 | AUS | F | 5 | 1245 | 18.60 | 0.66 |
| 6 | AUS | F | 6 | 609 | 8.80 | 0.42 |

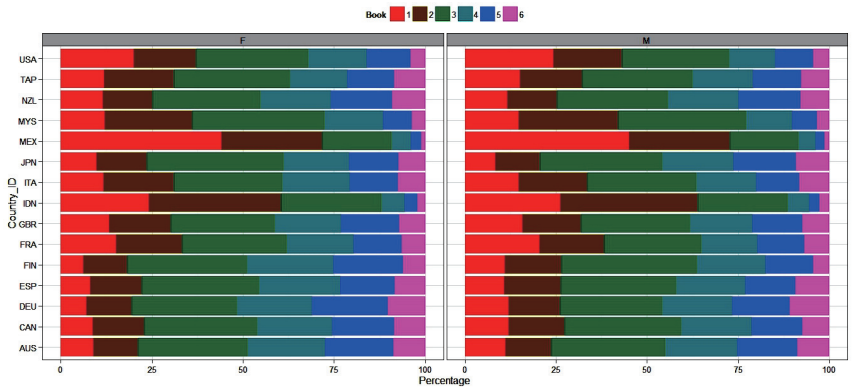


**圖 1**　臺灣與其它十四個國家男、女學生家中藏書量的次數分配圖

國家，墨西哥（MEX）不論男、女學生在這個部分的人數明顯少於其它十四個國家，此外墨西哥的家中藏書量只有 0 到 10 本書（1）明顯多於其它十四個國家。

以男、女學生與家中藏書量計算臺灣與其它十四個國家的平均數學成績，使用 pisa.mean.pv( ) 來進行運算，其中 pvlabel 是指成績變項，在此輸入 MATH 是代表整體數學成績，export 是指是否要將結果輸出為.csv 檔，以 head( ) 來檢視前六行結果，結果如表 4。並利用 plot( ) 繪製該結果，其中 na.omit( ) 是處理資料表格中的遺漏值，sort 是指是否要依平均數來排列，結果如圖 2，程式碼見附錄四。從圖 2 可知，臺灣（TAP）在每個家中藏書量的選項上，平均數學成績幾乎都高於其它十四個國家，除了在家中藏書量選項是 1 時低於日本，另外只有在家中藏書量有 500 本書以上（6）的部分，臺灣女學生平均數學成績是高於男學生，至於其它家

表 4

**以學生性別與家中藏書量計算臺灣與其它十四個國家的前六筆平均數學成績**

| | Country_ID | Gender | Book | Freq | Mean | s. e. | SD | s. e |
|---|---|---|---|---|---|---|---|---|
| 1 | AUS | F | 1 | 679 | 430.44 | 4.15 | 81.18 | 3.51 |
| 2 | AUS | F | 2 | 898 | 453.16 | 3.30 | 83.43 | 2.56 |
| 3 | AUS | F | 3 | 2039 | 491.86 | 2.44 | 84.91 | 2.54 |
| 4 | AUS | F | 4 | 1431 | 510.86 | 2.49 | 84.19 | 1.89 |
| 5 | AUS | F | 5 | 1245 | 541.48 | 3.41 | 88.38 | 2.39 |
| 6 | AUS | F | 6 | 609 | 543.47 | 4.25 | 93.77 | 3.31 |

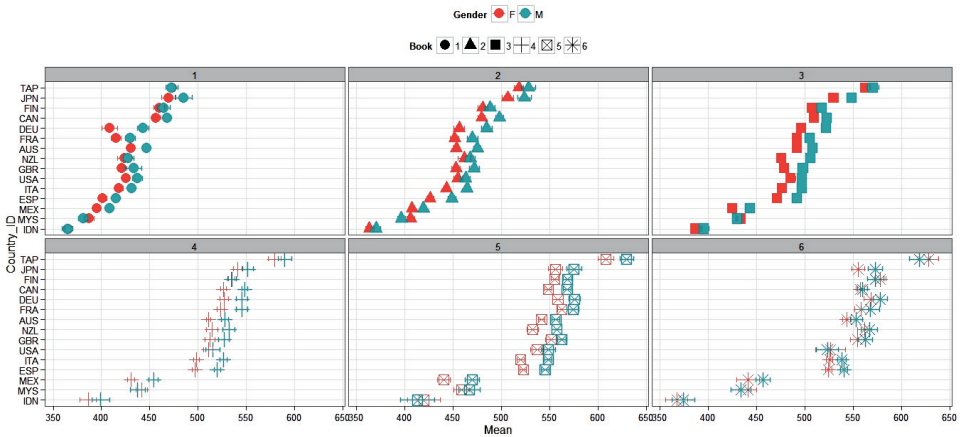


**圖 2**　以學生性別與家中藏書量繪製臺灣與其它十四個國家的平均數學成績圖

中藏書量的部分皆是男學生平均數學成績高於女學生。而印尼（IDN）在每個家中藏書量的選項上，平均數學成績都低於其它十四個國家，其中只有在家中藏書量有 201-500 本書（5）的部分，印尼女學生平均數學成績高於男學生，至於其它家中藏書量的部分皆是男學生平均數學成績高於女學生。

「迴歸」一直是作預測的主要模式，2005 年發表在美國〈教育社會學〉期刊的論文中就使用了 OLS 的迴歸（ordinary least square regression）、依變項為二分名義變項的邏輯迴歸（binary logistic regression）、依變項為多分名義變項的邏輯迴歸（multinomial logistic regression）等各種統計方

法。在教育社會學研究中，依變項大多是名義變項或次序變項，特別是在大型資料庫的資料，因此在進行資料庫研究的分析時，邏輯迴歸或對數線性模式等類似的分析方法就變成相當重要（黃鴻文，2006）。因此本研究運用 intsvy 套件進行過去研究中最常使用的迴歸分析。

在進行臺灣與其它十四個國家以男、女學生與家中藏書量、母親教育程度、家中擁有物對數學成績的預測之前，可在 R 先檢測資料是否滿足複迴歸模式的基本假設。接著將學生性別從因子轉換成數值，把學生性別的原本選項從女生是 1、男生是 2，改成女生是 0、男生是 1。以 pisa.reg.pv( ) 進行運算，其中 pvlabel 是指成績變項，在此輸入 MATH 代表整體數學成績，export 是指是否要把結果輸出為.csv 檔，以 head( ) 檢視前六行結果，其中 R-squared 在 intsvy 套件中，原先設定的範圍是 0 到 100，本研究將各國 R-squared 結果各自再除以 100，結果如表 5。以 plot( ) 繪製該結果，其中 se 是指是否需要加上標準誤，結果如圖 3，程式碼見附錄四。從圖 3 可知，在家中藏書量的 Estiamte 方面，法國（FRA）最高，印尼（IDN）最低。在學生性別的 Estiamte 方面，只有馬來西亞（MYS）是負向，其它十四個國家皆是正向。在母親教育程度的 Estiamte 方面，法國（FRA）最高，臺灣（TAP）最低。在家中擁有物的 Estiamte 方面，臺灣（TAP）最高，芬蘭（FIN）最低。在 R-squared 的 Estiamte 方面，家中藏書量、學生性別、母親教育程度與家中擁有物對數學成績的解釋力以法國（FRA）最高，印尼（IDN）最低。

以地圖方式呈現迴歸分析結果前，首先從迴歸分析結果中取出各變項的估計值，因此載入 plyr 套件，以 as.matrix[ ] 將原本是列表（list）型態的資料框轉換成矩陣（matrix），結果如表 6。再分別將學生性別、家中藏書量、母親教育程度、家中擁有物的估計值取出，以 data.frame( ) 把這四個矩陣型態的資料框建立成列表，並對這四個資料框內的各變項重新命名。接著利用 merge( ) 將學生性別、家中藏書量、母親教育程度、家中擁有物這四個資料框合併在同一個資料框內，以 c( ) 新增國家變項，將重新命名的臺灣與其它十四個國家名稱輸入，以便繪製微地圖時可與 R 的其它兩個套件作連結。以 str( ) 檢視新資料框內的資料結構後，結果如表 7，使用 as.factor( )、as.numeric( ) 將國家的變項型態從字串轉換成因子，將學生性別估計值、家中藏書量估計值、母親教育程度估計值、家中擁有物估計值從因子轉換成數值，以 str( ) 重新確認新資料框內的資料結構，結果如表 8。

表 5

**以學生性別、家中藏書量、母親教育程度、家中擁有物預測臺灣與其它十四個國家數學成績的前四個國家**

| $ AUS | | | |
|---|---|---|---|
| | Estimate | Std. Error | t value |
| (Intercept) | 332.78 | 5.80 | 57.36 |
| Gender | 15.81 | 2.70 | 5.86 |
| Book | 17.70 | 0.73 | 24.36 |
| Mother | 11.43 | 0.93 | 12.34 |
| Possessions | 5.38 | 0.57 | 9.50 |
| R-squared | 0.16 | 0.78 | 19.88 |
| $ CAN | | | |
| | Estimate | Std. Error | t value |
| (Intercept) | 374.98 | 9.31 | 40.30 |
| Gender | 16.03 | 1.92 | 8.36 |
| Book | 17.07 | 0.78 | 21.81 |
| Mother | 8.74 | 1.62 | 5.40 |
| Possessions | 3.70 | 0.51 | 7.32 |
| R-squared | 0.12 | 0.75 | 16.40 |
| $ DEU | | | |
| | Estimate | Std. Error | t value |
| (Intercept) | 339.39 | 14.68 | 23.12 |
| Gender | 23.59 | 2.93 | 8.05 |
| Book | 21.92 | 1.61 | 13.65 |
| Mother | 14.66 | 1.83 | 8.02 |
| Possessions | 3.77 | 1.30 | 2.89 |
| R-squared | 0.22 | 1.49 | 14.66 |
| $ ESP | | | |
| | Estimate | Std. Error | t value |
| (Intercept) | 314.16 | 5.66 | 55.52 |
| Gender | 20.12 | 2.15 | 9.38 |
| Book | 19.93 | 0.75 | 26.44 |
| Mother | 9.97 | 1.13 | 8.82 |
| Possessions | 5.11 | 0.58 | 8.87 |
| R-squared | 0.22 | 0.97 | 23.09 |

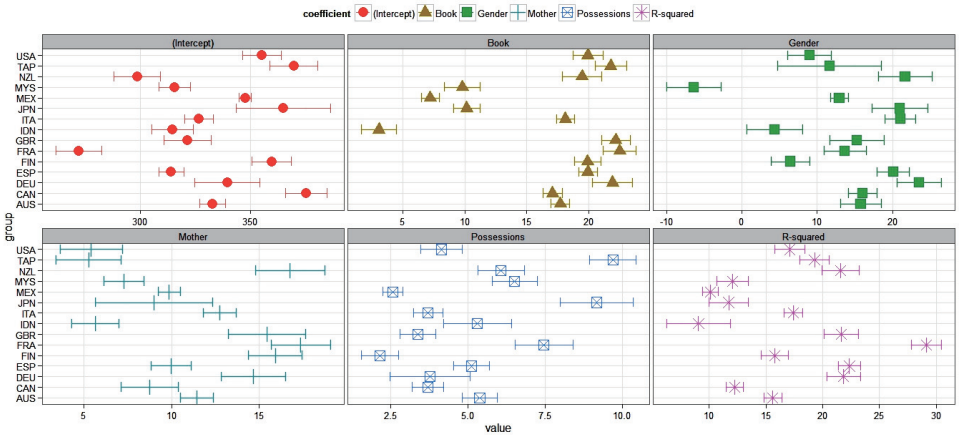


**圖 3** 以學生性別、家中藏書量、母親教育程度、家中擁有物預測臺灣與其它十四個國家的數學成績圖

表 6

**轉換資料框型態後迴歸分析結果的矩陣內容**

| | . id | Estimate | Std. Error | t value |
|---|---|---|---|---|
| [1, ] | “AUS” | “332.78” | “5.80” | “57.36” |
| [2, ] | “AUS” | “15.81” | “2.70” | “5.86” |
| [3, ] | “AUS” | “17.70” | “0.73” | “24.36” |
| [4, ] | “AUS” | “11.43” | “0.93” | “12.34” |
| [5, ] | “AUS” | “5.38” | “0.57” | “9.50” |
| [6, ] | “AUS” | “15.60” | “0.78” | “19.88” |

表 7

**新資料框的資料結構**

| | | | |
|---|---|---|---|
| ‘data.frame’ | : | 15 obs. of 6 variables: | |
| $ id | : | Factor w/ 15 levels | “AUS” “CAN” “DEU”,..: 1 2 3… |
| $ GenderEstimate | : | Factor w/ 15 levels | “-6.40”, “4.34”,..: 9 10 15 11 3… |
| $ BookEstimate | : | Factor w/ 15 levels | “3.09” , “7.20”,..: 6 5 13 11 9 15… |
| $ MotherEstimate | : | Factor w/ 15 levels | “5.28”, “5.43”,..: 9 5 11 8 13 15… |
| $ PossessionEstimate | : | Factor w/ 15 levels | “2.16”, “2.57”,..: 10 4 6 8 1 13 3… |
| $ COUNTRY | : | chr | “Australia” “Canada” “Germany” “Spain” … |

表 8

**新資料框轉換變項型態的資料結構**

| | | | | | | |
|---|---|---|---|---|---|---|
| 'data.frame' | : | 15 obs. of 6 variables: | | | | |
| $ id | : | Factor w/ 15 levels | "AUS" "CAN" "DEU",..: 1 2 3… | | | |
| $ GenderEstimate | : | num | 15.81 | 16.03 | 23.59 | 20.12 | 6.43 … |
| $ BookEstimate | : | num | 17.7 | 17.1 | 21.9 | 19.9 | 19.9 … |
| $ MotherEstimate | : | num | 11.43 | 8.74 | 14.66 | 9.97 | 15.9 … |
| $ PossessionEstimate | : | num | 5.38 | 3.7 | 3.77 | 5.11 | 2.16 … |
| $ COUNTRY | : | Factor w/ 15 levels | "Australia", "Canada",..: 1 2 5 12 … | | | | |

為了在地圖中呈現以學生性別估計值前面 25%、中間 50%、後面 25% 分類臺灣與其它十四個國家的情形，使用 quantile( ) 計算學生性別估計值 25% 與 75% 的數值，其數值分別是 10.310 與 20.575。利用 with( ) 在新資料框中新增 Genderlevel 變項，輸入學生性別估計值的最低數值、25% 數值、75% 數值、最高數值作為分類基準，命名各數值區間的名稱，分別為後面 25%、中間 50%、前面 25%，再輸入新資料框名稱檢視各資料內容，結果如表 9，程式碼見附錄五。

最後以地圖的方式呈現迴歸分析的結果，首先安裝與載入 maptools 與 RColorBrewer 套件。利用 readShapePoly( ) 讀取 world.shp，透過 str( ) 檢視資料結構，以 unique( ) 查看在 world.shp 中的 COUNTRY 名稱，需要注意的是新資料框中的 COUNTRY 必須要與 world.shp 中的 COUNTRY 的名稱相同，才能合併這兩個檔案的資料。合併 world.shp 與新資料框使用 merge( )，其中 by.y 是以兩個檔案都有的 COUNTRY 作為連結。colorRampPalette( ) 被用來調節顏色漸層以表示數值的高低。並使用 spplot( ) 分別將迴歸分析中臺灣與其它十四個國家的學生性別估計值、家中藏書量估計值、母親教育程度估計值、家中擁有物估計值以地圖的方式呈現，結果如圖 4、圖 5、圖 6、圖 7，程式碼見附錄六。

從圖 4 可知，性別估計值的前面 25% 國家分別是德國（DEU）、紐西蘭（NZL）、義大利（ITA）、日本（JPN），後面 25% 國家分別是馬來西亞（MYS）、印尼（IDN）、芬蘭（FIN）、美國（USA）。從圖 5 可知，法國（FRA）的顏色最深，即表示該國的家中藏書量估計值最高，印尼（IDN）的顏色最淺，這表示該國的家中藏書量估計值最低，相較於

表 9

**新資料框的資料內容**

| | id | Gender Estimate | Book Estimate | Mother Estimate | Possessions Estimate | COUNTRY | Gender level |
|---|---|---|---|---|---|---|---|
| 1 | AUS | 15.81 | 17.70 | 11.43 | 5.38 | Australia | 中間 50% |
| 2 | CAN | 16.03 | 17.07 | 8.74 | 3.70 | Canada | 中間 50% |
| 3 | DEU | 23.59 | 21.92 | 14.66 | 3.77 | Germany | 前面 25% |
| 4 | ESP | 20.12 | 19.93 | 9.97 | 5.11 | Spain | 中間 50% |
| 5 | FIN | 6.43 | 19.90 | 15.90 | 2.16 | Finland | 後面 25% |
| 6 | FRA | 13.70 | 22.49 | 17.36 | 7.46 | France | 中間 50% |
| 7 | GBR | 15.32 | 22.18 | 15.43 | 3.38 | United Kingdom | 中間 50% |
| 8 | IDN | 4.34 | 3.09 | 5.66 | 5.31 | Indonesia | 後面 25% |
| 9 | ITA | 21.05 | 18.11 | 12.75 | 3.71 | Italy | 前面 25% |
| 10 | JPN | 21.03 | 10.16 | 9.00 | 9.17 | Japan | 前面 25% |
| 11 | MEX | 12.95 | 7.20 | 9.87 | 2.57 | Mexico | 中間 50% |
| 12 | MYS | -6.40 | 9.80 | 7.28 | 6.51 | Malaysia | 後面 25% |
| 13 | NZL | 21.70 | 19.45 | 16.75 | 6.07 | New Zealand | 前面 25% |
| 14 | TAP | 11.64 | 21.78 | 5.28 | 9.69 | Taiwan | 中間 50% |
| 15 | USA | 8.98 | 19.92 | 5.43 | 4.15 | United States of America | 後面 25% |

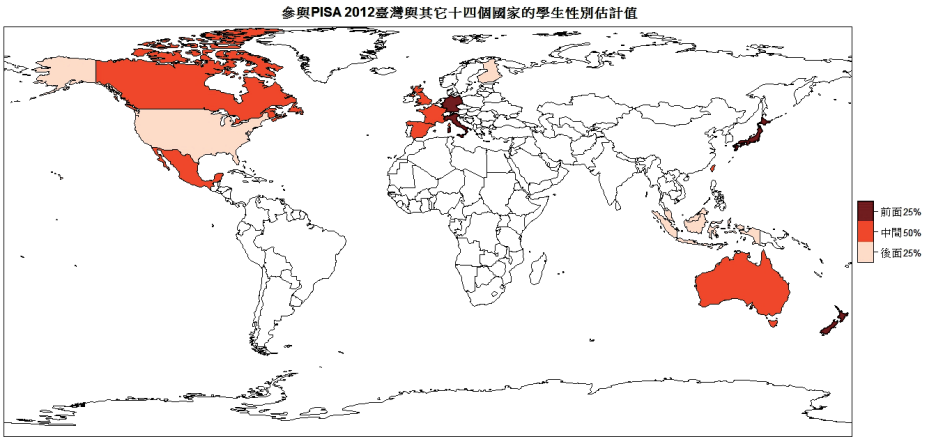


**圖 4**　參與 PISA 2012 臺灣與其它十四個國家的學生性別估計值

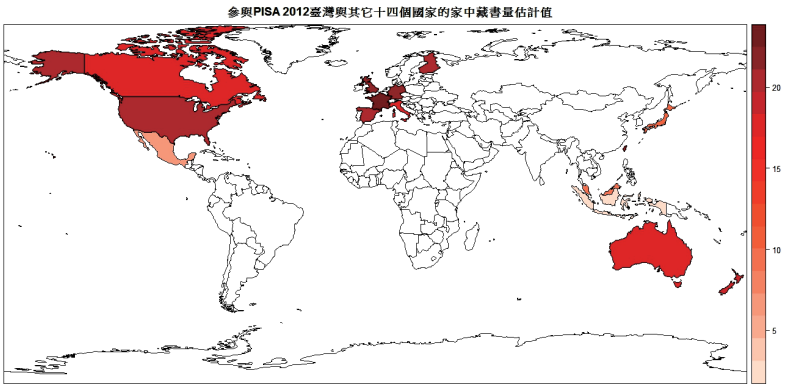


**圖 5**　參與 PISA 2012 臺灣與其它十四個國家的家中藏書量估計值

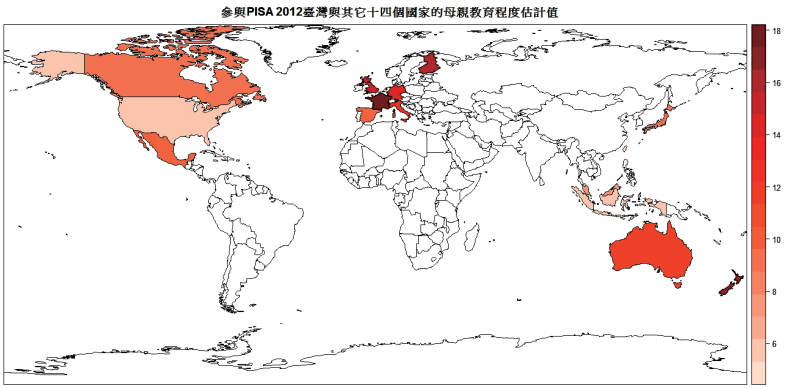


**圖 6**　參與 PISA 2012 臺灣與其它十四個國家的母親教育程度估計值

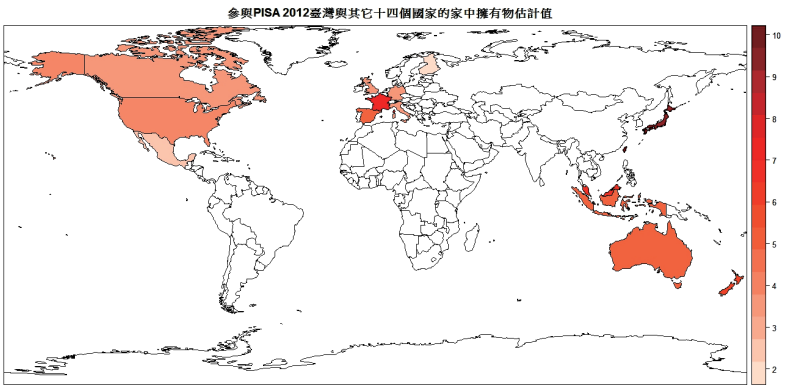


**圖 7**　參與 PISA 2012 臺灣與其它十四個國家的家中擁有物估計值

大部分的歐美與紐澳國家，亞洲國家的顏色幾乎呈現淺色。從圖 6 可知，法國（FRA）的顏色最深，即表示該國的母親教育程度估計值最高，臺灣（TAP）的顏色最淺，這表示該國的母親教育程度估計值最低，相較於其它地區的國家，歐洲國家的顏色多呈現深色。從圖 7 可知，臺灣（TAP）的顏色最深，即表示該國的家中擁有物估計值最高，芬蘭（FIN）顏色最淺，這表示該國的家中擁有物估計值最低，相較於其它地區的國家，臺灣與日本的顏色多呈現深色。

同時從圖 5、圖 6、圖 7 可知，北美國家的家中藏書量估計值的顏色比起其母親教育程度估計值與家中擁有物估計值的顏色較為深，這表示北美國家的家中藏書量每增加一個單位，數學成績就會增加一分的情況優於其母親教育程度與家中擁有物每增加一個單位，數學成績就會增加一分的情況。而歐洲國家的家中藏書量估計值與母親教育程度估計值的顏色比起其家中擁有物估計值的顏色較為深，即歐洲國家的家中藏書量估計值與母親教育程度每增加一個單位，數學成績就會增加一分的情況優於其家中擁有物每增加一個單位，數學成績就會增加一分的情況。在亞洲國家方面，臺灣與日本的家中擁有物估計值比起其家中藏書量估計值與母親教育程度估計值的顏色較為深，這表示臺灣與日本的家中擁有物每增加一個單位，數學成績就會增加一分的情況優於其家中藏書量估計值與母親教育程度每增加一個單位，數學成績就會增加一分的情況，而其它亞洲國家的家中藏書量估計值、母親教育程度估計值、家中擁有物估計值的顏色皆較為淺，即其它亞洲國家的家中藏書量、母親教育程度、家中擁有物每增加一個單位，數學成績就會增加一分的情況較不優於臺灣與日本的家中藏書量、母親教育程度、家中擁有物每增加一個單位，數學成績就會增加一分的情況。紐澳地區的家中藏書量估計值的顏色比起其母親教育程度估計值、家中擁有物估計值的顏色較為深，這表示紐澳地區的家中藏書量每增加一個單位，數學成績就會增加一分的情況優於其母親教育程度與家中擁有物每增加一個單位，數學成績就會增加一分的情況。

# 結論與建議

本研究的主旨在示範如何運用開放軟體 R 管理與分析大型教育資料庫的檔案。案例為，包括本國在內共十五個國家，十五歲學生數學成就表現與多個影響該表現的相關因素（PISA 2012）。範例分析，除了使用 R 的基本功能外，特別嵌入 Caro 與 Biecek（2014）針對處理大型教育資料庫發展的 intsvy 套件並結合 maptools 等相關套件將分析結果呈現在地圖上以利比較。此外提供數據與用來分析數據的電腦程式碼給其他人，才能達到對科學主張作評判的基本標準「複現」（Peng, 2011）。因此本文中提及的檔案管理及資料分析過程皆可使用提供在附錄的 R 程式碼複製。

從本研究運用 intsvy 套件進行分析的範例中可知，intsvy 套件處理大型教育資料庫數據的功能可節省研究者的時間，像是選擇所需的學生、家長、學校、教師教學的變項與國家，只需一個函數直接合併數據導入到 R，所有成績似真值與加權都已預設在內等功能。intsvy 套件分析與繪製大型教育資料庫數據的功能可供研究者以最短的時間進行多國比較，像是次數表、平均數、迴歸分析等分析，只需一個函數即可一次呈現各國的結果報表與圖形。雖然目前 intsvy 套件在繪製圖形方面只可使用在 PISA 與 PIAAC 這兩個資料庫所提供的數據檔，但 R 的使用者可自行開發套件的擴充，因此建議未來大型教育資料庫的相關研究，可將 intsvy 套件與 R 的其它分析（例如：結構方程模型、多層次分析等）、繪製的套件作結合，使研究結果能呈現得更加完整。

Summary

# USING R TO ANALYZE INTERNATIONAL LARGE-SCALE EDUCATIONAL ASSESSMENT DATA

## INTRODUCTION

With the increasing popularity of international large-scale assessments (e.g., TIMSS, PIRLS, PISA, and PIAAC), research on cross-country comparisons of educational achievement has also increased. Currently, researchers mostly use commercial statistical software to analyze educational data. However, stable and reliable replication is an important standard to judge the validity of results. The standard of reproducibility requires that the data and the computer code used to analyze the data must be accessible to other researchers (Peng, 2011). Researchers who do not have access to the same commercial software cannot easily replicate the results others have reported in publications. Therefore, computational findings often do not cumulate reliably to offer valid guidelines for educational policy making.

For open source platform, Ihaka and Gentleman (1996) developed R: A language and environment for statistical computing and graphics. R is highly extensible and has become very popular in statistical research community. Caro and Biecek (2014) created an R package “intsvy”, which is designed for importing, merging, analyzing, and graphing data from international assessment studies (TIMSS, PIRLS, PISA, and PIAAC). The package not only allows users to import data directly from selected student, parent, school, teacher variables, and countries into R, but also provides users with functions for routine statistical analysis such as computing descriptive summaries and regression modeling. In addition, the “intsvy” package considers the complex sample design (i.e., replicate weights) and rotated test forms (i.e., plausible achievement values) to calculate point estimates and standard errors.

This paper demonstrates how to use the free, open-source R computing environment (R Core Team, 2015) to manage and analyze international large-scale educational assessment data. The illustrative data example consists of mathematics achievement and covariates (student gender, number of books at

home, mother's highest schooling and possessions) for 15-year-old students, from 15 countries, including Taiwan (PISA 2012). The "intsvy" package is applied to manage the data, and the "maptools" package (Bivand & Lewin-Koh, 2015) is used to link numerical summaries to geographical boundaries of the countries examined. R script to perform analysis of data from PISA detailed in the paper is provided to facilitate reproducibility.

## METHOD

The data example included mathematics achievement and covariates for 15-year-old students from participating countries in PISA 2012. The countries selected are Taiwan (TAP), Australia (AUS), Canada (CAN), Germany (DEU), Spain (ESP), Finland (FIN), France (FRA), United Kingdom (GBR), Indonesia (IDN), Italy (ITA), Japan (JPN), Mexico (MEX), Malaysia (MYS), New Zealand (NZL), and United States of America (USA). The covariates selected were student gender, number of books at home, mother's highest schooling, and possessions.

The merge functions and analysis functions of the 'intsvy' package are used in the following steps:

### 1. Select and Merge Data

Functions "pisa.select.merge" is used to select and merge data from PISA; all achievement and weight variables are selected automatically.

### 2. Frequency Table and Graph of Frequency Table

Function "pisa.table" is used to provide a frequency table for a categorical variable, including percentages and standard errors. Also, function "plot.intsvy.table" is used to present this frequency table graphically. However, it is worth noting that function "plot.intsvy.table" is only used for data from PISA and PIAAC.

### 3. Mean of Variable and Graph of Means in Groups

Function "pisa.mean" is used to produce the mean of selected variable and its standard error, but the selected variable does not include plausible values. Function "pisa.mean.pv" is used to calculate the mean achievement and its standard error using five plausible values. In addition, function

"plot.intsvy.mean" is used to show these means graphically. However, it is worth noting that function "plot.intsvy.mean" is only used for data from PISA and PIAAC.

**4. Regression Analysis with Plausible Values and Graphs of Regression Models in Groups**

Function "pisa.reg.pv" is used to implement linear regression analysis with plausible values and replicate weights. Also, function "plot.intsvy.reg" is used to present the result of linear regression analysis graphically. Currently, the function "plot.intsvy.reg" can only be used for data from PISA and PIAAC.

Finally, we used the "maptools" package (Bivand & Lewin-Koh, 2015) to represent numerical summaries of linear regression analysis on the geographical boundaries of the countries.

## RESULTS

This paper successfully demonstrates how to use an open-source, computing environment, R, and its contributed packages such as the "intsvy" package and the "maptools" package to manage and analyze data directly from international assessment studies (TIMSS, PIRLS, PISA, and PIAAC) to achieve aims in reproducible research.

## CONCLUSIONS AND SUGGESTIONS

**1. Conclusions**

R can effectively manage and analyze international large-scale educational assessment data. Most importantly, replication is the ultimate standard to judge scientific claims (Peng, 2011), and R certainly conforms to the standard of reproducibility. Because R can provide other researchers with the data and the computer code used to analyze the data, researchers can easily replicate results that others have reported in publications through R. Therefore, reliably cumulating computational findings can offer policymakers valid guidelines.

**2. Suggestions**

(1) The “intsvy” package can be used to explore cross-country comparisons.

(2) The “intsvy” package can be used to match other statistical computing (e.g., structural equation modeling and hierarchical linear modeling) and graphics packages.

# 參考文獻

柯華葳、詹益綾、丘嘉慧（2013）。**臺灣四年級學生閱讀素養－PIRLS 2011 報告**。2015 年 4 月 9 日，取自：http://lrn.ncu.edu.tw/Teacher%20web/hwawei/Project/PIRLS%202011%E5%AE%8C%E6%95%B4%E5%A0%B1%E5%91%8A.pdf

[Ko, H. W., Chan, Y.-L., & Chiu, C.-H. (2013). *PIRLS 2011 National report: Reading literacy study of fourth grade Taiwanese students*. Retrieved April 9, 2015, from http://lrn.ncu.edu.tw/Teacher%20web/hwawei/Project/PIRLS%202011%E5%AE%8C%E6%95%B4%E5%A0%B1%E5%91%8A.pdf]

陳靜怡（2012）。**利用基因組富集分析方法分析基因微陣列資料－以十字花科黑腐病菌感染阿拉伯芥為模型**（未出版之碩士論文）。亞洲大學，臺中市。

[Chen, C.-Y. (2012). *Investigation of microarray data using gene set enrichment analysis—arabidopsis thaliana infected with xanthomonas campestris pv. campestris* (Unpublished master's thesis). Asia University, Taichung.]

麥馨月（2004）。**臺灣教育研究的發展與趨勢－以教育研究所學位論文為研究對象**（未出版之碩士論文）。國立高雄師範大學，高雄市。

[Mai, H. Y. (2004). *The status and trends of educational research in Taiwan--subjects on thesis and dissertation of graduate school of education* (Unpublished master's thesis). National Kaohsiung Normal University, Kaohsiung.]

黃鴻文（2006）。2005 年美國〈教育社會學〉期刊之回顧與啟示。**中等教育**，**57**(4)，142-145。

[Huang, H.-W. (2006). The recollection and implications on American's "Sociology of Education" in 2005. *Secondary Education*, *57*(4), 142-145.]

葉信伶、鄒惠貞、江宏哲、江博煌、劉德明（2013）。利用階層式分群法點熱圖以瞭解工業污染地區的死亡風險因子。**醫療資訊雜誌**，**22**(2)，1-14。

[Yeh, H.-L., Tsou, H.-C., Chiang, H.-C., Chiang, P.-H., & Liou, D.-M. (2013). The use of heat map with hierarchical clustering in mortality risk to identify "Hot-Spots" risk factors in polluted industrial areas. *The Journal of Taiwan Association for Medical Informatics*, *22*(2), 1-14.]

臺灣 PISA 國家研究中心（2014）。**臺灣 PISA 2012 精簡報告**。2014 年 12 月 1 日，取自：http://pisa.nutn.edu.tw/download/data/TaiwanPISA2012ShortReport.PDF

[Taiwan PISA National Center. (2014). *Taiwan PISA 2012 short report*. Retrieved December 1, 2014, from http://pisa.nutn.edu.tw/download/data/TaiwanPISA2012ShortReport.PDF]

臺灣 PISA 國家研究中心（2015）。**關於 PISA。計畫概述**。2015 年 3 月 26 日，取自：http://pisa.nutn.edu.tw/pisa_tw.htm

[Taiwan PISA National Center. (2015). *About PISA. Overview*. Retrieved March 26, 2015, from http://pisa.nutn.edu.tw/pisa_tw.htm]

鄭中平、許清芳（2015）。**R 在行為科學之應用**。臺北市：雙葉書廊。

[Cheng, C.-P., & Sheu, C.-F. (2015). *R for behavior research*. Taipei, Taiwan: Yeh Yeh Book Gallery.]

Benton, T. (2014). Using meta-regression to explore moderating effects in surveys of international achievement. *Practical Assessment, Research & Evaluation*, *19*(3), 1-9.

Bivand, R., & Lewin-Koh, N. (2015). maptools: Tools for Reading and Handling Spatial Objects. R package version 0.8-34. Retrieved from http://CRAN.R-project.org/package=maptools

Caro, D., & Biecek, P. (2014). intsvy: International Assessment Data Manager. R package version 1.6. Retrieved from http://CRAN.R-project.org/package=intsvy

Chambers, J. M. (2009). *Software for data analysis: Programming with R* (2nd Ed.). New York, NY: Springer.

Fox, J. (2009). Aspects of the social organization and trajectory of the R project. *The R Journal*, *1/2*, 5-13.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S.,…Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.1-R80.16.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314.

Liou, P.-Y., & Hung, Y.-C. (2014). Statistical techniques utilized in analyzing PISA and TIMSS data in science education from 1996 to 2013: A methodological review. *International Journal of Science and Mathematics Education*. doi: 10.1007/s10763-014-9558-5

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). *TIMSS 2007 International science report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Organisation for Economic Co-operation and Development. (2010). *Programme for the international assessment of adult competencies.* Retrieved March 25, 2015, from http://www.oecd.org/

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ripley, B. D. (2001). The R project in statistical computing. *MSOR Connections*, *1*(1), 23-25.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International largescale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142-151.

Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education, 54,* 627-650.

Versteeg, R., Richardson, A., & Rowe, T. (2006). Web-accessible scientific workflow system for performance monitoring. *Environmental Science and Technology*, *40*(8), 2692-2698.

Vidal, D., Thormann, M., & Pons, M. (2005). LINGO, An efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *Journal of Chemical Information and Modeling*, *45*(2), 386-393.

# 附錄一

```
#讀取資料並檢視資料結構
dta<-read.csv("PISA2012TAP.csv",header=TRUE,na.string=".")
str(dta)

#轉換變項型態
dta$ST28Q01<-as.numeric(dta$ST28Q01)

#以迴歸分析預測家中藏書量對數學成績的影響，並產出報表
dta.lm <- lm(PV1MATH ~ ST28Q01,data=dta)
summary(dta.lm)
```

# 附錄二

```
#安裝與載入 intsvy 套件
install.packages("intsvy")
library(intsvy)

#選擇與合併數據
dta <- pisa.select.merge(folder="C:/Users/ANAN/Desktop",
student.file="INT_STU12_DEC03.sav",
student= c("ST04Q01","ST28Q01","ST13Q01","ST26Q01","ST26Q02",
"ST26Q03","ST26Q04","ST26Q05","ST26Q06","ST26Q07","ST26Q08",
"ST26Q09","ST26Q10","ST26Q11","ST26Q12","ST26Q13","ST26Q14"),
countries = c("AUS","CAN","DEU","ESP","FIN","FRA","GBR","IDN",
"ITA","JPN","MEX","MYS","NZL","TAP","USA"))

#檢視資料結構，結果如表 1
str(dta)
```

# 附錄三

```
#轉換變項型態
dta$CNT<-as.factor(dta$CNT)
for(i in 4:20) {
dta[,i] <- as.numeric(dta[,i])
}

#轉換變項選項與名稱
dta$Country_ID<-dta$CNT

dta$ST04Q01<-abs(dta$ST04Q01-1)
dta$Gender<-dta$ST04Q01
dta$Gender<-factor(dta$Gender, levels=c(0,1), labels=c("F","M"))

dta$Book<-dta$ST28Q01

dta$ST13Q01<-abs(dta$ST13Q01-6)
dta$Mother<-dta$ST13Q01

for(i in 6:19) {
dta[,i] <- abs(dta[,i]-2)
}
dta$Possessions<-dta$ST26Q01+dta$ST26Q02+dta$ST26Q03+dta$ST26Q04
+dta$ST26Q05+dta$ST26Q06+dta$ST26Q07+dta$ST26Q08+dta$ST26Q09
+dta$ST26Q10+dta$ST26Q11+dta$ST26Q12+dta$ST26Q13+dta$ST26Q14

#再次檢視資料結構，結果如表 2
str(dta)
```

# 附錄四

```
#計算臺灣與其它十四個國家男、女學生家中藏書量的前六筆次數分配，並繪製臺灣與其它十四個國家的結果，結果如表 3 與圖 1
head(ptableCB <- pisa.table(variable="Book", by=c("Country_ID", "Gender"), data=dta))
plot(na.omit(ptableCB), stacked=TRUE)

#以學生性別與家中藏書量計算臺灣與其它十四個國家的前六筆平均數學成績，並繪製臺灣與其它十四個國家的結果，結果如表 4 與圖 2
head(pmeansMCGB <- pisa.mean.pv(pvlabel="MATH", by=c("Country_ID", "Gender", "Book"), data=dta, export=FALSE))
plot(na.omit(pmeansMCGB), sort=TRUE)
```

# 附錄四（續）

```
#以學生性別、家中藏書量、母親教育程度、家中擁有物預測臺灣與其它十四
個國家數學成績的前四個國家，並繪製臺灣與其它十四個國家的結果，結果
如表 5 與圖 3
dta$Gender<-as.numeric(dta$Gender)
dta$Gender<-abs(dta$Gender-1)
head(rmodelMGBMP <- pisa.reg.pv(pvlabel="MATH",
x=c("Gender","Book","Mother","Possessions"), by = "Country_ID", data=dta,
export=FALSE))
plot(rmodelMGBMP, se=TRUE)
```

# 附錄五

```
#下載 plyr 套件
library(plyr)

#將迴歸分析結果的資料框型態轉換成矩陣
rmodelMGBMPmatrix<-as.matrix(ldply(rmodelMGBMP))

#檢視矩陣內的資料，結果如表 6
head(rmodelMGBMPmatrix)

#取出矩陣內臺灣與其它十四個國家的學生性別估計值、家中藏書量估計值、
母親教育程度估計值與家中擁有物估計值，並將其資料框型態轉換成列表，
且重新命名變項名稱
dataGM<-rmodelMGBMPmatrix[c(2,8,14,20,26,32,38,44,50,56,62,68,74,80,86),1:2]
dataG<-data.frame(dataGM)
colnames(dataG)<-c("id","GenderEstimate")

dataBM<-rmodelMGBMPmatrix[c(3,9,15,21,27,33,39,45,51,57,63,69,75,81,87),1:2]
dataB<-data.frame(dataBM)
colnames(dataB)<-c("id","BookEstimate")

dataMM<-rmodelMGBMPmatrix[c(4,10,16,22,28,34,40,46,52,58,64,70,76,82,88),1:2]
dataM<-data.frame(dataMM)
colnames(dataM)<-c("id","MotherEstimate")

dataPM<-rmodelMGBMPmatrix[c(5,11,17,23,29,35,41,47,53,59,65,71,77,83,89),1:2]
dataP<-data.frame(dataPM)
colnames(dataP)<-c("id","PossessionsEstimate")
```

# 附錄五（續）

```
#合併臺灣與其它十四個國家的學生性別估計值、家中藏書量估計值、
母親教育程度估計值與家中擁有物估計值，並存入新資料框
dataGB<-merge(dataG,dataB,by="id",all=TRUE)
dataGBM<-merge(dataGB,dataM,by="id",all=TRUE)
dataGBMP<-merge(dataGBM,dataP,by="id",all=TRUE)

#在新資料框中新增重新命名的國家變項
dataGBMP$COUNTRY<-c("Australia","Canada","Germany","Spain","Finland",
"France","United Kingdom","Indonesia","Italy","Japan","Mexico","Malaysia",
"New Zealand","Taiwan","United States of America")

#檢視資料結構，結果如表 7
str(dataGBMP)

#轉換變項型態
dataGBMP$COUNTRY<-as.factor(dataGBMP$COUNTRY)
dataGBMP$GenderEstimate<-as.numeric(as.character(dataGBMP$GenderEstimate))
dataGBMP$BookEstimate<-as.numeric(as.character(dataGBMP$BookEstimate))
dataGBMP$MotherEstimate<-as.numeric(as.character(dataGBMP$MotherEstimate))
dataGBMP$PossessionsEstimate<-as.numeric
(as.character(dataGBMP$PossessionsEstimate))

#再次檢視資料結構，結果如表 8
str(dataGBMP)

#計算學生性別估計值的前面 25%與後面 25%的數值
quantile(dataGBMP$GenderEstimate,probs = c(0.25,0.75))

#建立新增學生性別估計值程度變項
dataGBMP$Genderlevel <- with(dataGBMP, cut(GenderEstimate,
breaks=c(-6.41,10.310,20.575,23.59), labels=
c("後面 25%","中間 50%","前面 25%")))

#檢視新資料框內的資料，結果如表 9
dataGBMP
```

# 附錄六

```
#安裝與載入 maptools 與 RColorBrewer 套件
install.packages("maptools")
install.packages("RColorBrewer")
library("maptools")
library("RColorBrewer")

#讀取 world.shp 與檢視其資料結構
world.shp<-readShapePoly("./world/world.shp")
str(world.shp)

#檢視在 world.shp 中的國家名稱
unique(world.shp$COUNTRY)

#合併 world.shp 與新資料框
rwrldGBMP<-merge(world.shp,dataGBMP,by.y="COUNTRY",all.x=TRUE)

#設定顏色漸層
cols <- colorRampPalette(brewer.pal(6, "Reds"))(53)

#繪製臺灣與其它十四個國家的學生性別估計值，結果如圖 4
spplot(rwrldGBMP, "Genderlevel", col.regions=cols,main="參與 PISA 2012 臺灣
與其它十四個國家的學生性別估計值")

#繪製臺灣與其它十四個國家的家中藏書量估計值，結果如圖 5
spplot(rwrldGBMP, "BookEstimate", col.regions=cols,main="參與 PISA 2012 臺
灣與其它十四個國家的家中藏書量估計值")

#繪製臺灣與其它十四個國家的母親教育程度估計值，結果如圖 6
spplot(rwrldGBMP, "MotherEstimate", col.regions=cols,main="參與 PISA 2012
臺灣與其它十四個國家的母親教育程度估計值")

#繪製臺灣與其它十四個國家的家中擁有物估計值，結果如圖 7
spplot(rwrldGBMP, "PossessionsEstimate", col.regions=cols,main="參與 PISA
2012 臺灣與其它十四個國家的家中擁有物估計值")
```