教育數據拾遺：如何使用不完整的資訊

駐波士頓臺北經濟文化辦事處文化組

政策制定者和分析師在使用「輸入調整法」評比大學院校時，必須盡可能掌握真實的數據資料。在現實世界中，教育方面的數據常常並不完整、也不容易使用。即便有強大的分析工具或方法，資料蒐集往往比分析本身更為困難，錯誤的資料就可能導致錯誤的結論。

有鑑於此，參與 HCM「成功的脈絡」這一系列研究的資深研究人員列出了使用數據的一些重點：

廣泛蒐集資料

有許多種不同來源的資料可用於「輸入調整法」，數據的多寡則視蒐集的時間和來源而定。以下是可以蒐集資料的方向：

- 12 年國教期間的成績單和其他記錄（分數、是否有免費午餐、課表、操行等）；
- 大學申請資料（有考慮但並未就讀的學校）；
- 財務援助申請資料（收入、財產）；
- 高等教育記錄（轉學學分抵免，結果變項）；
- ACT 與 SAT 測驗記錄（分數，背景調查問題）；
- 就業記錄 （結果變項，上大學前的受雇記錄）；
- 學生調查 （參見〈正確的大學院校調查〉一文）；
- 全國學生情報交換中心（私校/跨州轉學）。

讓遺漏的資料發揮作用

將以上資料合併之後，許多學生的記錄可能有部分遺漏。分析師應該建立一個獨立的變項，說明每個變項是否有遺漏，並將這些變項也做分析。某些資料的欠缺本身就具有意義，例如，沒有參加 SAT 或 ACT 測驗的學生，可能原本就不打算就讀四年制大學；未曾填寫 FAFSA（聯邦學生補助免費申請）的學生更可能來自高收入環境。

避免遺漏

分析師應盡可能使用所有可獲得的資訊，否則選擇性的過濾資訊可能導致上百種不同的預測結果。如果某些資訊最終無助於預測結果、或不容易取得、或難以解釋，之後隨時可以將這些資料拿掉，但任何可能有用的資訊都值得一試。

修正偏離值或使用對數

在收入這個變項，一個 Facebook 員工的收入可以是一群大學畢業生平均收入的兩倍。有一種替代方法是將這些極端值排除或設定上限，但這樣可能漏掉重要資訊（畢竟這所學校的確出了至少一個成功的工程師！）。更好的辦法

是以收入的對數來代替收入本身的數字，這樣讓100萬到10萬美元的差距與10萬到1萬美元的差距相仿，而不是差到10倍。其他會出現巨大偏離值的變項也應採類似的處理方式。

追蹤多個世代

儘可能採用不只一年的入學學生資訊，這樣有兩個潛在的優點，首先能增加資訊的數量，讓統計效果更好；其次，這樣統計出來的結果可能更接近一般真實情況，而非某一年度學生的特殊情形。但若使用太多年份的資料，統計結果可能無法反映當下的現況。

譯稿人：魏瑀嫻

參考資料:10/22/2012 HCM 策士公司研究報告終篇