**從 pd.get\_dummies 到 sklearn.linear\_model.LinearRegression**

還記得剛接觸Pandas 套件的 get\_dummies時，當下彷彿天降甘霖，解決了在資料清洗當中，如何有效處置類別型變數(Categorical Variable)的困難，想當初最一開始敲代碼時，還一個一個用 if…elif….else 去新增虛擬變數(dummy Variable)。

然而現實總是不若想像中美好，如果將下表(A)使用 pd.get\_dummies 時，會得到表B：

表A

|  |  |  |
| --- | --- | --- |
| weight | age | Gender |
| 40 | 10 | M |
| 72 | 15 | F |
| 51 | 13 | F |
| 80 | 20 | M |

表B

|  |  |  |  |
| --- | --- | --- | --- |
| weight | age | Gender\_M | Gender\_F |
| 40 | 10 | 1 | 0 |
| 72 | 15 | 0 | 1 |
| 51 | 13 | 0 | 1 |
| 80 | 20 | 1 | 0 |

然而根據經濟學的教法，在建立線性模型OLS時，若對類別型變數取Dummy Variable，且該類別型變數有k個去重覆值，Dummy Variable的個數將會是k-1個。

因此，模型理論上應該是：

而實際上，若直接使用pd.get\_dummies則會變成one-hot encoding的型態：

所幸，pd.get\_dummies提供一個參數 drop\_first 可以個別將不同類別型變數get\_dummies後的第一個欄位刪除，然而，根據粗略瀏覽網路上各個python機器學習、python資料分析等等的教學或文件，甚少會使用drop\_first，而是直接以one-hot encoding的型態去訓練模型。

在說明我的困惑前，先稍微解釋為何經濟學的計量模型會對於Dummy Variable的個數那麼講究。

首先，由於多數線性模型會放入常數項(亦即截距項)，如此一來常數項、Gender\_M和Gender\_F便會產生完全共線性，其中一個變數會被其他變數完全解釋