行政院國家科學委員會專題研究計畫　成果報告

可合併性的分析方法在列聯表資料和迴歸模型間的可能整合

計畫類別：個別型計畫
計畫編號：NSC93-2118-M-004-005-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立政治大學統計學系

計畫主持人：江振東

報告類型：精簡報告

處理方式：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中　華　民　國 94 年 11 月 3 日

行政院國家科學委員會專題研究計畫成果報告

可合併性的分析方法在列聯表資料和回歸模型間的可能整合

On Analyzing Collapsibility in Contingency Tables and Regression Models

計畫編號：NSC 93-2118-M-004-005

執行期限：93 年 8 月 1 日至 94 年 7 月 31 日

主持人：江振東 國立政治大學統計系

一、計畫中文摘要

可合併性(collapsibility)的問題在列聯表資料分析中已被廣泛討論，類似的問題在迴歸分析中也有過探討，然而兩者間的可能互通性卻似乎不曾有過討論。Clogg, Petkova 及 Shihadeh(1992)介紹一種新的檢定方式來分析迴歸問題中的可合併性，並且進一步應用到列聯表資料中。在此計畫中我們就可合併性這個概念在迴歸分析及列聯表資料分析中的關聯性作瞭解，並試圖沿用列聯表資料分析中的幾種分析可合併性的檢定方式來處理迴歸模型中相類似問題的可行性。此外透過廣義線性模型，我們得到一種探討可合併性的檢定方式。模擬結果顯示，此一檢定方式有相當不錯的整體表現。

關鍵詞：可合併性、列聯表、迴歸分析、廣義線性模型

Abstract

Issues on collapsibility in contingency tables have been well-documented. Similar issues which were stated as full model, reduced model setting in linear regression problems can also be found in the literature. Although the issues are apparently related, they were treated independently in the two areas. Clogg, Petkova, and Shihadeh (1992) was the only exception which introduced a new procedure for assessing collapsibility in regression problems, and also discussed the possible extensions to the contingency tables setting. In this study, we tried to sort out the differences among the various treatments on collapsibility in contingency tables. Using generalized linear models, we provided a statistical test that can be used to see if the idea of collapsibility can be applied to a given data set. Based on simulation studies, we found that the result is quite promising.

Keywords: collapsibility, contingency tables, regression analysis, generalized linear models

二、計畫緣由與目的

這個計畫基本上可以視為我在前一年度所執行的國科會計畫「邊際列聯表的適用性探討」的一個延續。由於在針對列聯表的可合併性(collapsibility)作相關文獻搜尋和回顧的過程中，經由閱讀了由 Clogg, Petkova 及 Shihadeh 於 1992 發表於 Journal of Educational Statistics 的文章“Statistical Methods for Anaiyzing Collapsibility in Regression Models”後所啟發的想法。該文章的主要著眼點如下：

假定就一組樣本資料我們分別配適下述兩個迴歸模型：

(A)簡化模型(reduced model)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_{\mathbf{R}} + \boldsymbol{\varepsilon}_{\mathbf{R}}$$

(B)完全模型(full model)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_{\mathbf{R}} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_{\mathbf{F}}$$

其中 **X**, **Z** 分別為 $n\times p$ 及 $n\times q$ 的矩陣。**X** 係由我們所感到興趣的自變數(variables of interest)所構成，而 **Z** 則是由次要自變數(covariates)所組成。在一般傳統的迴歸分析書籍中，焦點可能是擺在檢定 $\boldsymbol{\gamma}$ 是否為 **0**，或者是探討 model under-fitting 時所可能造成的影響。不過該文章所要探討的重點則是在於比較係數 $\boldsymbol{\beta}_{\mathbf{R}}$ 與 $\boldsymbol{\beta}_{\mathbf{F}}$ 的異同。他們想要瞭解在忽略次要自變數的情形下，主要變數 **X** 所對應的係數 $\boldsymbol{\beta}$ 是否依然維持不變。

其實這類問題的探討，在列聯表(contingency tables)資料中我們常會碰到。由於一個列聯表的維數(dimension)如果太大，在做資料分析時往往會造成許多的不便，同時也會增加模型解釋上的困難度。因此一種直觀的思維方式便是設法藉由忽略其中某些次要因子(secondary factors)來降低維數，並且利用這些維數較低的邊際列聯表(marginal tables)來作必要的分析。這個方法如果可行的話，問題就簡化多了。然而 Simpson's paradox 告訴我們，這樣的作法未必完全能夠採行。就某些結構的列聯表而言，主要因子在原始列聯表中所呈現的關係在邊際列聯表中或許可以維持不變，但就絕大多數的情形，則完全是兩回事。這一類問題的探討，在列聯表資料分析中常用的名詞就稱作可合併性。然而可合併性這個性質的討論，並非僅只適用在列聯表的資料分析上。前述線性迴歸模型結構中所提及的簡化模型(A)、完全模型(B)，儘管其自變數未必如列聯表資料一樣為類別型變數，但其模型結構的寫法和問題所要探討的主旨其實和列聯表資料是一致的。雖然針對模型(A)(B)而言，瞭解增加自變數 **Z** 是否會對我們在預測 **y** 時所有幫助，絕對是一個重要課題，事實上相關文獻及書籍也以此為主軸，並有詳盡的解釋和說明。不過如果一個研究的主要目的只是想要瞭解 **X** 與 **y** 之間的關係，而且如果次要自變數 **Z** 的存在與否並不會造成主要自變數 **X** 對 **y** 的解釋有任何改變(亦即係數 $\boldsymbol{\beta}$ 維持不變)的話，那麼我們又何必一定要採用完全模型(B)。如果直接採用簡化模型(A)來解釋 **X** 與 **y** 之間的關係，問題也就簡化許多。Clogg, Petkova 及 Shihadeh(1992)舉了許多例子來說明這類問題的重要性，同時沿用列聯表資料分析中可合併性的基本概念，針對前述迴歸模型中 $\boldsymbol{\beta}_{\mathbf{R}} = \boldsymbol{\beta}_{\mathbf{F}}$ 的現象，稱呼次要自變數 **Z** 可以被合併掉(collapsible)，並就此一問題提出一個新的統計檢定方法，探討該檢定所具有的性質，同時也舉出實例來作說明。此外該文中也就該檢定方法在列聯表資料中應用的可行性，作了一些衍生的探討和說明。

由於目前文獻中關於列聯表資料的可合併性有不同的定義，依照條件要求的深淺程度以及適用情況的不同，由簡而繁可以區分為簡易合併性(collapsibility)，嚴格合併性(strict collapsibility)及強固合併性(strong collapsibility)(參見 Whittemore(1978), Shapiro(1982), Ducharme and Lepage(1986))，然而按照 Clogg, Petkova 及 Shihadeh(1992)就迴歸問題中的定義看來，他們的方法在列聯表資料上的應用範圍似乎僅與簡易合併性的結構有關。因此這也提供我們一個新的思考模式，我們想要瞭解其他兩種條件較強的合併性，是否也可以應用於線性迴歸的問題，而其實際的義涵又是如何，這可能會有助於我們來處理實際問題。此外由於列聯表資料分析多半採用的對數線性模型(log-linear model)及一般的線性迴歸模型都屬於廣義線性模型(generalized linear model)的範疇，因此我們也想瞭解進一步將前述的

這些討論推廣至廣義線性模型的可行性。

三、計畫結果與討論

由於我們的主要目的是希望就列聯表資料及迴歸分析模型的可合併性，作整合探討，因此廣義線性模型無疑是最適當的出發點。

假定就一組樣本資料，我們分別配適下述的兩個廣義線性模型：

(A) $\boldsymbol{\eta}_R = \mathbf{X}_1\boldsymbol{\beta}_R$，其中 $\boldsymbol{\eta}_R = g(\boldsymbol{\mu}_R)$，$\mathbf{X}_1$ 是一個 $n \times p$ 的矩陣。

(B) $\boldsymbol{\eta}_F = \mathbf{X}\boldsymbol{\beta} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_F \\ \boldsymbol{\gamma}\end{pmatrix} = \mathbf{X}_1\boldsymbol{\beta}_F + \mathbf{X}_2\boldsymbol{\gamma}$，其中 $\boldsymbol{\eta}_F = g(\boldsymbol{\mu}_F)$。

假設模型(B)是個 true model。在此前提之下，依據 Clogg et. al. (1992)，我們的問題可以轉換成探討 $\boldsymbol{\beta}_R = \boldsymbol{\beta}_F$ 是否成立。

藉由廣義線性模型的理論推導 (McCullagh and Nelder(1989))，我們知道在模型(B)中，$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$，其中 $\mathbf{W} = \hat{\text{cov}}(\mathbf{Y})$，$\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{W}^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_F)$，$\hat{\boldsymbol{\mu}}_F = g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}})$。因此 $\hat{\boldsymbol{\beta}}_F = (\mathbf{I}_p, \mathbf{0})\hat{\boldsymbol{\beta}}$。

同理，模型(A)中的參數也可透過相同方式求得：$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{W}_R\mathbf{z}_R$，其中 $\mathbf{W}_R = \hat{\text{cov}}_R(\mathbf{Y})$，$\mathbf{z}_R = \mathbf{X}_1\hat{\boldsymbol{\beta}}_R + \mathbf{W}_R^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_R)$，$\hat{\boldsymbol{\mu}}_R = g^{-1}(\mathbf{X}_1\hat{\boldsymbol{\beta}}_R)$。定義 $\boldsymbol{\delta} = \boldsymbol{\beta}_R - \boldsymbol{\beta}_F$。由於我們想要檢定的是 $\boldsymbol{\delta}$ 是否等於 $\mathbf{0}$，因此 $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F$。此外

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\delta}}) &= \text{cov}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F) \\ &= \text{cov}(\hat{\boldsymbol{\beta}}_R) + \text{cov}(\hat{\boldsymbol{\beta}}_F) - 2\,\text{cov}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}_F)\end{aligned}$$

這其中的三個分量可以進一步表示如下：

就模型(B)而言，

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}) &= \text{cov}((\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}) \\ &\approx (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\end{aligned}$$

因此

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}_F) &= \text{cov}((\mathbf{I}, \mathbf{0})\hat{\boldsymbol{\beta}}) \\ &\approx (\mathbf{I}, \mathbf{0})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix}\end{aligned}$$

同理，就模型(A)而言，

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}_R) &\approx (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\text{cov}(\mathbf{Y})\mathbf{X}_1(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &= (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}(\mathbf{I}, \mathbf{0})\mathbf{X}'\,\text{cov}(\mathbf{Y})\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix}\mathbf{X}(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\end{aligned}$$

再者，

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}) &= \text{cov}((\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{W}_R\mathbf{z}_R, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}) \\ &\approx (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}(\mathbf{I}, \mathbf{0})\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\end{aligned}$$

因此

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}_F) &= \text{cov}(\hat{\boldsymbol{\beta}}_R, (\mathbf{I}, \mathbf{0})\hat{\boldsymbol{\beta}}) \\ &\approx (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}(\mathbf{I}, \mathbf{0})\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix}\end{aligned}$$

綜合上述幾項結果，我們可以得知

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\delta}}) &= \text{cov}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F) \\ &\approx (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}(\mathbf{I}, \mathbf{0})\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix}(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &\quad +(\mathbf{I}, \mathbf{0})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix} \\ &\quad -2(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}(\mathbf{I}, \mathbf{0})\mathbf{X}'\,\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix}\end{aligned}$$

由於我們假定模型(B)為真，所以我們可以利用 cov(**Y**)來估計 **W**，也就是說 $\hat{\text{cov}}(\mathbf{Y}) = \mathbf{W}$。因此

$$\begin{aligned}\hat{\text{cov}}(\hat{\boldsymbol{\delta}}) &= (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}(\mathbf{I}, \mathbf{0})\mathbf{X}'\mathbf{W}\mathbf{X}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix}(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &\quad +(\mathbf{I}, \mathbf{0})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix} - 2(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &= (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\,\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_F)(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &\quad +\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_F) - 2(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\end{aligned}$$

這是 $\hat{\text{cov}}(\hat{\boldsymbol{\delta}})$ 的通式。不過在虛無假設成立的情況下，我們可以進一步的加以簡化。由於此時 $\text{cov}(\hat{\boldsymbol{\beta}}_F) = \text{cov}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}_F)$，因而 $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_F)$ 與 $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}_F)$ 都是 $\text{cov}(\hat{\boldsymbol{\beta}}_F)$ 的一個 consistent estimator。然而 $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}_F) = (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}$ 較 $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_F)$ 更 efficient，所以在虛無假設為真的前提下，我們可以嘗試以 $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\beta}}_F)$ 來取代 $\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_F)$。我們因此可以得到

$$\begin{aligned}\hat{\text{cov}}(\hat{\boldsymbol{\delta}}) &= (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}\,\hat{\text{cov}}(\hat{\boldsymbol{\beta}}_R,\hat{\boldsymbol{\beta}}_F)(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &\quad + \hat{\text{cov}}(\hat{\boldsymbol{\beta}}_R,\hat{\boldsymbol{\beta}}_F) - 2(\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &= (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-3} - (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1} \\ &= (\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-1}((\mathbf{X}_1'\mathbf{W}_R\mathbf{X}_1)^{-2} - \mathbf{I})\end{aligned}$$

透過 $\hat{\boldsymbol{\delta}}$ 和 $\hat{\text{cov}}(\hat{\boldsymbol{\delta}})$，我們便能夠針對 $\boldsymbol{\delta}$ 是否為 0 的問題作檢定，同時就可合併性的問題作探討。

上述計算 $\hat{\text{cov}}(\hat{\boldsymbol{\delta}})$ 的公式，其結構相當單純，不過藉由模擬實驗我們發現它的表現相當好。我們試著與 Clogg et. al(1992)所曾經提及的幾種估計 $\text{cov}(\hat{\boldsymbol{\delta}})$ 的方式作比較，結果大體上都是相當正面的。

四、計畫成果自評

這個計畫基本上是延續去年的計畫而來，去年我們僅僅針對列聯表的資料作探討，而今年我們則有機會再就可合併性的問題，在迴歸分析與列聯表的架構下，重新作回顧。過去的文獻中，針對可合併性問題的論述，都是分開來作討論。Clogg et.al.(1992)似乎是唯一的例外。它提供了不同的思維，也就是引進廣義線性模式，透過廣義線性模式來綜合討論可合併性的問題。由於相關的文獻並不多，因此我們只能透過有限的資源來作發掘。在這裡，我們所得到的結果與 Clogg et.al.(1992)最主要的差異在於使用不同的估計量來估計 $\text{cov}(\hat{\boldsymbol{\delta}})$。雖然這似乎只是個小小的突破，不過藉由模擬實驗的方式，我們發現我們所使用的估計量的表現要比 Clogg et.al.(1992)所建議的幾種方式，都要來的好。這或許應該是執行此計畫的最大收穫了。當然我們也仍有些未盡之處。就列聯表資料的可合併性而言，有所謂的簡易合併性、嚴格合併性及強固合併性，這三類合併性如何透過廣義線性模式來進一步作詮釋，我們並沒有太具體的突破。未來我們還會找機會再針對這一部份繼續作了解。

五、參考文獻

1.Asmussen, S. and D. Edwards (1981). Collapsibility and Response Variable in Contingency Tables. *Biometrika*, 70, 567-578.

2.Chow, G.C. (1960). Tests of Equality between Sets of Coefficients in Two Linear Regressions. *Econometrika*, 28,591-605.

3.Clogg, C. C., E. Petkova, and E. S. Shihadeh (1992). Statistical Methods for Analyzing Collapsibility in Regression Models. *Journal of Educational Statistics*, 17, 51-74.

4.Ducharme, G. R. and Y. Lepage (1986). Testing Collapsibility in Contingency Tables. *Journal of the Royal Statistical Society, Series B*, 48, 197-205.

5.McCullagh, P. and J. A. Nelder. (1989). *Generalized Linear Models*, Chapman and Hall, London.

6.Shapiro, S. H. (1982). Collapsing Contingency Tables: A geometric Approach. *American Statistician*, 36, 43-46.

7.Whittemore, A, (1978). Collapsibility of Multidimensional Contingency Tables. *Journal of the Royal Statistical Society, Series B*, 40,328-340.