

大學指考落點預測與實證分析

任眉眉¹ 陳日昇² 林家立¹

¹ 國立成功大學統計學系

² 國立成功大學計算機與網路中心

摘 要

落點分析是大學指定考試分發入學前，考生選填志願的重要輔助工具，該分析主要建立在「前後兩年學生對校系基礎評價不變」的前提下。我們 [1] 於九十四學年度首次以統計理論探討該前提假設是否成立，並應用九十二和九十三學年度聯分會及大考中心公佈的相關資料，進行實證研究。在前文中，我們直接假設所有考科成績均具有常態分佈，且變異數均相同，進而應用連續兩學年度各考科成績的中位數、各校系之指考科目加權權重、前一學年度各校系最低錄取加權總分等資料，建立一個落點預測模型。本文主要提出一個驗證各考科成績是否具有常態分佈的統計推論，並進一步估計各考科成績常態分佈的平均數與變異數，再根據這些估計值修正前文中的落點預測模型。另外，本文亦以多變量分析理論，探討前文中採用線性迴歸模型作為落點預測模型的合理性，並依據預測區間的原理，首次建立評估落點分析準確度的方法。最後，我們針對九十三到九十六學年的資料進行實證分析，說明我們所建立的預測模型具有相當高的準確度，並首度完成一個多年大學指考分發結果的具體研究。

關鍵詞：最低錄取加權總分，常態分佈，多變量分析，線性迴歸模型，預測區間，準確度。

美國數學協會分類索引：主要 62J05；次要 62P25。

1. 前言

爲了導引考生在大學選填志願時能符合自己的性向，使各大學能夠依據自己的特色招收適才學生，遂於民國八十八年六月確定所謂的「考招分離，多元入學」方案，並於民國九十一學年度首次實施。大學多元入學方案歷經多年修正，目前各校系均要求考生需參加學測檢定，且各校系採計考試科目的加權權重逐年修正。因應這種變革，考生如何運用歷年來聯合分發會（簡稱聯分會）所公佈的各校系最低加權錄取總分，及當年度大考中心所公佈的各單一考科成績累計表（含頂標、前標、均標、後標和底標等五標），在一千六百多個校系中，依序填選出符合自己志願，又能夠避免高分低就的校系？

每年大學指考成績寄發以後，坊間都會推出很多免費或收費的落點分析，但是多數做過落點分析的考生都覺得效果不顯著。如何評估各種落點分析的準確度？這也是一個有趣的研究問題。基於推廣統計應用的理念，我們 [1] 曾於九十四學年度首次以統計理論與方法，提出一個落點預測模型。進而寫成一個有效的電腦落點分析軟體和操作手冊，置放於成大統計系網頁上 [4]，提供考生及家長一個免費且高效率的選填志願參考。

基本上，落點分析是建立在「前後兩年學生對校系基礎評價不變」的前提假設下，落點分析效果不顯著的主要原因，可能是各校系前後兩個年度採計科目的權重有些修正，也可能是考題難易度有所不同。[1] 中，我們首次以統計理論探討該前提假設是否成立，並應用九十二和九十三學年度聯分會及大考中心公佈的資料，進行實證研究。文中，我們直接假設所有考科成績均具有常態分佈 (normal distribution)，且變異數 (variance) 均相同。進而應用連續兩學年度，大考中心公佈的各考科成績的中位數 (median)、各校系之指考科目加權權重，及聯分會公佈的各校系最低錄取加權總分等資料，建立一組評價指標，應用這些評價指標的相關係數，探討該前提假設是否成立。接著，依據這些評價指標的線性迴歸模型，建立一個落點預測模型。

前文中，我們直接假設所有考科成績均具有常態分佈，因此本文首先探討如何驗證各考科成績具有常態分佈。我們將應用大考中心所發佈的各單一考科考生成績累計表，提出驗證各考科成績具有常態分佈的統計方法，並進一步估計各考科成績分佈的平均數 (mean) 與變異數，再根據這些估計值，修正前文中的評價指標及落點預測模型。另外，本文亦以多變量分析理論，探討 [1] 中採用線性迴歸模型作為落點預測模型的合理性，並依據預測區間的原理，首次建立評估落點分析準確度的方法。最

後，我們針對九十三到九十六學年的資料進行實證分析，說明我們所建立的預測模型具有相當高的準確度。

在第2節中，我們主要探討落點預測中的統計模型與統計推論，首先定義一些符號及各校系的評價指標，然後說明如何應用校系評價指標來驗證前提假設成立。再探討如何驗證各考科成績是否服從常態分佈，及常態分佈中平均數與變異數的估計方法。並在常態分佈下，探討採用線性迴歸模型作為落點預測的合理性。最後，在線性迴歸模型下，依據預測區間的原理，建立評估落點分析準確度的方法。在第3節中，我們將針對九十三到九十六學年度大考中心與聯分會所公佈的資料，先剔除新一學年度新增及加考術科的校系後，再依所有校系一起或分成19個學群來分別討論，配合第2章的統計理論逐步進行實證分析。

2 統計模型與分析

聯分會所公佈的各校系最低錄取加權總分，是依據各校系採計考試科目的權重計算，採計科目愈多或權重愈大，均會導致最低錄取總分有被灌水的可能性。前文建議使用加權平均，作為比較前後兩年學生對校系基礎評價是否改變的評價指標，並假設各考科成績分佈為常態分佈。由於不同年度各考科題目難易度不同，且同年度各考科題目難易度也不盡相同，本節主要探討如何應用大考中心所發佈的各單一考科成績累計表（含頂標、前標、均標、後標和底標，分別表示各考科第88%、75%、50%、25%和12%考生的成績），驗證前文中各考科成績服從常態分佈的假設，並進一步估計該常態分佈的平均數與變異數。其次，我們利用估計的平均數與變異數，調整因題目難易度及採計考試科目權重不同所造成的差異，提出一個修正的評價指標及落點預測模型，據此準確地預估各校系新年度最低錄取加權總分。

2.1 符號定義

首先，令 $i = 1, \dots, 9$ 分別代表各考試科目（國文、英文、數甲、數乙、歷史、地理、物理、化學和生物）。針對某固定年度某固定考科，令 X 代表考科成績，則 $E(X) = \mu$ 及 $Var(Y) = \sigma^2$ 分別代表該考科成績分佈的平均數與變異數。若連續考慮 J 年，再令 μ_{ij} 與 σ_{ij}^2 分別代表年度 j 考科 i 成績分佈的平均數與變異數，

$i = 1, 2, \dots, 9, j = 1, \dots, J$ 。在不失通常性下，我們假設不同年度所有考科成績間互相統計獨立 (statistically independent)。

針對某固定校系，令 X_{ij} 代表該校系年度 j 最低錄取總分考生考科 i 之成績， w_{ij} 代表該校系年度 j 考科 i 之加權權重， w_{ij} 的可能值為 0, 1.0, 1.25, 1.5, 1.75, 2.0，則

$$T_j = \sum_{i=1}^9 w_{ij} X_{ij}$$

代表年度 j 聯分會所公佈的該校系最低錄取加權總分。由於 T_j 的平均數為

$$\mu_{T_j} = E\left(\sum_{i=1}^9 w_{ij} X_{ij}\right) = \sum_{ik=1}^9 w_{ij} \mu_{ij},$$

表示加權計分方式不同，會導致錄取總分有被灌水的可能性，而且在所有考科成績間互相統計獨立的假設下，其變異數為

$$\sigma_{T_j}^2 = Var\left(\sum_{i=1}^9 w_{ij} X_{ij}\right) = \sum_{i=1}^9 w_{ij}^2 \sigma_{ij}^2.$$

註：我們沒有全部考生的成績，此獨立假設無法進一步驗證是否成立。

由於不同年度各考科題目難易度不同，且同年度各考科題目難易度也不盡相同，因此我們將不同年度該校系的最低錄取加權總分標準化後 (standardize)，再作為各校系的評價指標以供後續研究用，亦即

$$Z_j = \frac{T_j - \mu_{T_j}}{\sigma_{T_j}} = \frac{\sum_{i=1}^9 w_{ij} (X_{ij} - \mu_{ij})}{\sqrt{\sum_{i=1}^9 w_{ij}^2 \sigma_{ij}^2}}, \quad (1)$$

則我們有 $E(Z_j) = 0$ 且 $Var(Z_j) = 1, j = 1, 2, \dots, J$ 。實務應用上，由於大考中心並未公佈各考科的平均數與變異數，我們將於 2.3 節中提出 μ_{ij} 與 σ_{ij}^2 的估計方法。

2.2 前提假設的驗證及應用

為了提升落點分析的準確性，在落點分析預測前，需先檢驗前提假設「前後兩年學生對校系基礎評價不變」是否成立。本研究的預測模型適用於某一學群或學類，亦可以將全部校系一併考量，令 K 為該學群或學類中校系總數。我們採用 [1] 中的作法，在連續兩年放榜後，考量某校系連續兩年，標準化後的最低錄取加權指標

(Z_1, Z_2) 。透過觀察 (Z_{1k}, Z_{2k}) , $k = 1, \dots, K$ 的散佈圖 (scatter plot), 我們可以初步瞭解上述前提假設是否成立, 亦可進一步應用 Z_1 與 Z_2 的相關係數 (coefficient of correlation)

$$r_{Z_1, Z_2} = \frac{\sum_{k=1}^K (Z_{1k} - \bar{Z}_1)(Z_{2k} - \bar{Z}_2)}{\sqrt{\sum_{k=1}^K (Z_{1k} - \bar{Z}_1)^2 \sum_{k=1}^K (Z_{2k} - \bar{Z}_2)^2}},$$

來探討 Z_1 與 Z_2 的相關性, 其中 $\bar{Z}_j = \sum_{k=1}^K Z_{jk}/K$ 。若前後兩年評價指標 (Z_{1k}, Z_{2k}) , $k = 1, \dots, K$ 的線性相關程度高, 即代表前提假設「前後兩年學生對校系基礎評價不變」成立。

若我們進一步假設所有考科成績具有常態分佈, 經由2.4節中的理論探討, 我們即可合理地應用簡單線性迴歸模型 (simple linear regression model)

$$E(Z_{2k}|Z_{1k}) = \beta_0 + \beta_1 Z_{1k}, \quad k = 1, \dots, K, \quad (2)$$

透過對 β_0 和 β_1 的統計推論研究, 建立一個合理的落點分析的統計預測模型。有關線性迴歸模型分析理論, 請參考 [6]。事實上, 若各考科成績確實服從常態分佈, 則我們可應用多重線性迴歸模型 (multiple linear regression)

$$E(Z_{Jk}|Z_{1k}, \dots, Z_{(J-1)k}) = \beta_0 + \beta_1 Z_{1k} + \dots + \beta_{J-1} Z_{(J-1)k}, \quad k = 1, \dots, K. \quad (3)$$

透過對 $\beta_0, \beta_1, \dots, \beta_{J-1}$ 的研究, 利用歷年資料 (Z_1, \dots, Z_{J-1}) 建立一個 Z_J 的落點預測模型, 詳見2.4節。

有關如何驗證線性迴歸模型中的假設「各考科成績服從常態分佈」, 我們會在2.3節中提出一個新的統計方法, 並在第3節中, 針對九十三到九十六學年度的資料, 進行實證分析。

2.3 常態假設的驗證及參數估計

在2.2節中, 我們先假設所有考科成績具有常態分佈, 進而建立落點預測模型 (2)。本節我們將介紹如何應用大考中心所發佈的, 各單一考科考生成績累計表中的 n 個 p -百分位數 (p th percentile), 來驗證「各考科成績服從常態分佈」的假設, 並進一步估計各考科成績常態分佈的平均數與變異數。爲了簡單起見, 文中我們僅應用各單一

考科考生成績的五個 p -百分位數 (頂標、前標、均標、後標和底標) 來說明並進行資料分析。

2.3.1 常態假設的驗證

假設某考科成績具有常態分佈, 以 $X \sim N(\mu, \sigma^2)$ 表之, 則 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. 注意:

$$Z = \frac{X - \mu}{\sigma} \iff X = \mu + \sigma Z \quad (4)$$

令 x_p 表示該成績分佈的 p -百分位數, 則大考中心所發佈的各單一考科考生成績的五標 ($n = 5$, 頂標、前標、均標、後標和底標), 分別記作 x_{p_i} , 其中 $p_1 = 88$ 、 $p_2 = 75$ 、 $p_3 = 50$ 、 $p_4 = 25$ 、 $p_n = 12$ 。對應於標準常態分佈, 我們可以查表找出 z_{p_i} 值, 根據 (4) 式我們應該有

$$x_{p_i} = \mu + \sigma z_{p_i}, i = 1, 2, \dots, n.$$

因此, 如同 2.2 節中的作法, 我們可以先針對 (x_{p_i}, z_{p_i}) , $i = 1, 2, \dots, n$ 製作散佈圖。若呈現直線關係, 即可進一步應用簡單線性迴歸模型

$$E(X_{p_i} | z_{p_i}) = \mu + \sigma z_{p_i}, i = 1, 2, \dots, n$$

來估計該考科成績分佈的平均數與變異數, 並透過殘差分析 (residual analysis) 了解上述模型的配適度。如果僅用五個資料點 ($n=5$), 執行簡單線性迴歸分析比使用散佈圖具有說服力。由於 z_{p_i} 值對應於標準常態分配, 若殘差平方和相當小時, 即表示考科成績 X 確實服從常態分佈。

2.3.2 參數估計

若考科成績 X 服從常態分佈, 則其平均數與變異數的估計方法及性質如下: 考慮線性迴歸模型

$$X_{p_i} = \mu + \sigma z_{p_i} + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} N(0, \nu^2), \quad i = 1, 2, \dots, n. \quad (5)$$

根據線性迴歸分析理論 [6], 我們有模型中斜率 (slope) σ 及截距 (intercept) μ 的最小平方估計量 (least square estimator) 表示如下:

$$\begin{cases} \hat{\sigma} = \frac{\sum(z_{p_i} - \bar{z}_p)(x_{p_i} - \bar{x}_p)}{\sum(z_{p_i} - \bar{z}_p)^2} \\ \hat{\mu} = \bar{x}_p - \hat{\sigma} \bar{z}_p. \end{cases}$$

令 $e_i = X_{p_i} - \hat{X}_{p_i} = X_{p_i} - \hat{\mu} - \hat{\sigma}z_{p_i}$, $i = 1, \dots, n$ 且 $s^2 = \sum_{i=1}^n e_i^2/n - 2$, 則在模型 (5) 下, 我們有

$$\frac{\hat{\mu} - \mu}{s/\sqrt{n}} \sim t(n-2) \quad \text{且} \quad \frac{\hat{\sigma} - \sigma}{s/\sqrt{\sum_{i=1}^n (z_{p_i} - \bar{z}_p)^2}} \sim t(n-2)$$

其中 $t(r)$ 表示自由度為 r 的 t -分佈。查表可知各個 z_{p_i} 值, 當僅使用五標時, 依據對稱性, 我們有 $z_{p_1} = -z_{p_5}$, $z_{p_2} = -z_{p_4}$, $z_{p_3} = 0$ 且 $\sum z_{p_i} = 0$ 。進一步化簡後可得 μ 與 σ 的估計量

$$\begin{cases} \hat{\sigma} = \frac{\sum z_{p_i} x_{p_i}}{\sum z_{p_i}^2}, \\ \hat{\mu} = \bar{x}_p \end{cases} \quad (6)$$

2.4 常態假設與線性迴歸模型

本節主要探討在 2.2 節中所述, 假設所有考科成績均具有常態分佈下, 模型 (2) 成立的合理性。首先, 假設歷年來所有考科成績 X_{ij} 均服從常態分佈, 則 (1) 中的評價指標具有標準常態分佈, $Z_j \sim N(0, 1)$, $j = 1, \dots, J$ 。令 $\mathbf{Z} = (Z_J, Z_{J-1}, \dots, Z_1)'$, 則 \mathbf{Z} 服從多變量常態分佈, 亦即 $\mathbf{Z} \sim MN_J(\mathbf{0}, I)$, 其中期望值向量 $\mathbf{0} = (0, \dots, 0)'$ 為 J 維零向量, 共變異方陣 (covariance matrix) I_J 為 $J \times J$ 單位矩陣。

根據多變量分析理論, 我們知道若 $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})' \sim MN_p(\mu, \Sigma)$ 服從 p 維多變量常態分佈, 其中 $\mathbf{X}^{(1)}$ 為 q 維隨機向量, $\mathbf{X}^{(2)}$ 為 $p - q$ 維隨機向量, 同理令

$$\mu = (\mu^{(1)}, \mu^{(2)})' \quad \text{且} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \text{則給定 } \mathbf{X}^{(2)} = \mathbf{x}^{(2)} \text{ 之下, } \mathbf{X}^{(1)}|\mathbf{x}^{(2)} \text{ 仍然服從}$$

q 維多變量常態分佈, 其期望值及共變異方陣分別為

$$E(\mathbf{X}^{(1)}|\mathbf{x}^{(2)}) = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}^{(2)} - \mu^{(2)}) = \nu(\mathbf{x}^{(2)}), \quad (7)$$

$$E\{[\mathbf{X}^{(1)} - \nu(\mathbf{x}^{(2)})][\mathbf{X}^{(1)} - \nu(\mathbf{x}^{(2)})]'\mid\mathbf{x}^{(2)}\} = \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \quad (8)$$

有關多變量分析理論，詳細請參考 [5]。

若 $\mathbf{Z} = (\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)})' \sim MN_J(\mu, \Sigma)$ ，其中 $\Sigma = [\sigma_{ij}]$ ，令 $\mathbf{Z}^{(1)} = Z_J$ ， $\mathbf{Z}^{(2)} = (Z_{J-1}, \dots, Z_1)'$ ，(7) 式即多重線性迴歸模型 (3)。當 $J = 2$ 時，令 $\mu_j = E(Z_j)$ ， $\sigma_j^2 = \text{Var}(Z_j)$ ， ρ 為 Z_1 與 Z_2 的相關係數，則我們有

$$\begin{aligned} E(Z_2|z_1) &= \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(z_1 - \mu_1) \\ &= (\mu_2 - \rho \frac{\sigma_2}{\sigma_1} \mu_1) + \rho \frac{\sigma_2}{\sigma_1} z_1 \end{aligned}$$

此即為模型 (2)。特別是當 $\mu_j = 0, \forall j$ 時，(7) 式可進一步表示為

$$E(Z_J|z_1, \dots, z_{J-1}) = \Sigma_{12} \Sigma_{22}^{-1} (z_{J-1}, \dots, z_1)', \quad (9)$$

此結果即說明我們將在 2.5 節中，採用常數項為 0 的線性迴歸模型 (11)，做為落點預測模型的合理性。值得一提的是，在 $\rho_{ij} \equiv \rho, \forall i, j = 1, \dots, J$ 的特殊情形行下，我們可以推得 (8) 式與 (9) 式的精簡結果，證明過程如下：首先我們有

性質 1: 令 $A_{n \times n} = \begin{bmatrix} x & \rho & \cdots & \rho \\ \rho & x & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & x \end{bmatrix}$ ，則 A 的行列式為 $|A| = (x + (n-1)\rho)(x - \rho)^{n-1}$ 。

證明:

$$\begin{aligned} \begin{vmatrix} x & \rho & \cdots & \rho \\ \rho & x & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & x \end{vmatrix} &= (x + (n-1)\rho) \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \rho & x & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & x \end{vmatrix} \\ &= (x + (n-1)\rho) \begin{vmatrix} 1 & 0 & \cdots & 0 \\ \rho & x - \rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho & 0 & \cdots & x - \rho \end{vmatrix} \\ &= (x + (n-1)\rho)(x - \rho)^{n-1}, \text{ 故得證。} \end{aligned}$$

系1: 當 $\Sigma_{J \times J} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$ 時, 其行列式為 $|\Sigma| = (1 + (J - 1)\rho)(1 - \rho)^{J-1}$.

性質 2: 令 $I_{n \times n}$ 為單位矩陣, $E_{n \times n} = [e_{ij}]$ 為所有元素均為1的矩陣, 則

$$(pI + qE)^{-1} = \frac{1}{p} \left(I - \frac{q}{p + nq} E \right), \quad p + nq \neq 0.$$

證明: 因為 $E^2 = nE$, 我們有

$$\begin{aligned} (pI + qE) \frac{1}{p} \left(I - \frac{q}{p + nq} E \right) &= I + \frac{q}{p} E + \frac{q}{p + nq} E + \frac{q^2}{p(p + nq)} E^2 \\ &= I + \frac{q(p + nq) - pq - nq^2}{p(p + nq)} E \\ &= I. \end{aligned}$$

同理

$$\frac{1}{p} \left(I - \frac{q}{p + nq} E \right) (pI + qE) = I, \quad \text{故得證。}$$

系1中, $\Sigma_{J \times J} = (1 - \rho)I + \rho E$, 依據性質 2 我們有

系2: 系1中, $\Sigma_{J \times J}^{-1} = \frac{1}{(1 - \rho)} \left[I - \frac{\rho}{1 + (J - 1)\rho} E \right]$.

由於 Σ_{22} 可表為一個型如 $(1 - \rho)I + \rho E$ 的 $J - 1$ 階方陣, 依據系 2 我們有

系3: (8) 式中, $\Sigma_{22}^{-1} = \frac{1}{(1 - \rho)} \left[I - \frac{\rho}{1 + (J - 2)\rho} E \right]$.

同時, $\Sigma_{12} = (\rho, \cdots, \rho)$ 為一個 $J - 1$ 維的列向量, 直接計算可得

$$\Sigma_{12} \Sigma_{22}^{-1} = \frac{\rho}{1 + (J - 2)\rho} (1, 1, \cdots, 1),$$

代入 (8) 式及 (9) 式, 我們有

性質 3: $\forall J \geq 2$,

$$E(Z_J | z_1, \cdots, z_{J-1}) = \frac{\rho}{1 + (J - 2)\rho} (z_{J-1} + \cdots + z_1)$$

及

$$\Sigma_{11.2} = 1 - \frac{(J-1)\rho^2}{1+(J-2)\rho}.$$

2.5 預測模型的準確度

綜合言之, 根據歷年的資料, 我們可以依循下列步驟逐步進行統計資料分析。

步驟 1. 經由 2.3 節的估計方法, 先探討年度 j 考科 i 在簡單線性迴歸模型中, 截距 μ_{ij} 與斜率 σ_{ij} 的估計量 (6) 及其準確性。進一步將 (6) 代入 (1) 中得到年度 j 各校系的評價指標

$$Z_j = \frac{T_j - \sum_{i=1}^9 w_{ij} \hat{\mu}_{ij}}{\sqrt{\sum_{i=1}^9 w_{ij}^2 \hat{\sigma}_{ij}^2}}, \quad j = 1, 2, \dots, J. \quad (10)$$

步驟 2. 在建立落點預測模型前, 我們應用 2.2 節中所提出的方法, 針對 K 個校系連續兩年的資料 $(Z_{jk}, Z_{(j+1)k})$, $k = 1, \dots, K$, 進行驗證「前後兩年學生對校系的基礎評價不變」的前提假設是否成立。

步驟 3. 在前提假設成立下, 考慮簡單線性迴歸模型 (2), 探討虛無假設 $\beta_0 = 0$ 及 $\beta_1 = 1$ 是否成立, 透過相似於 2.3.2 節中的 t 分佈, 執行 t 檢定以便建立落點預測模型。

步驟 4. 若應用歷年資料檢定後, 虛無假設 $\beta_0 = 0$ 及 $\beta_1 = 1$ 均成立, 則在新年度的放榜前, 我們直接預估新年度的 $\hat{\beta}_0 = 0$ 且 $\hat{\beta}_1 = 1$ 亦成立。因此針對每個校系 k , 我們採用

$$\hat{Z}_{j+1} = \hat{\beta}_0 + \hat{\beta}_1 Z_j = Z_j$$

來預測新年度 $(j+1)$ 各校系的最低加權錄取總分 \hat{T}_{j+1} , *i.e.*

$$\frac{\hat{T}_{j+1} - \sum_{i=1}^9 w_{i(j+1)} \hat{\mu}_{i(j+1)}}{\sqrt{\sum_{i=1}^9 w_{i(j+1)}^2 \hat{\sigma}_{i(j+1)}^2}} = \frac{T_j - \sum_{i=1}^9 w_{ij} \hat{\mu}_{ij}}{\sqrt{\sum_{i=1}^9 w_{ij}^2 \hat{\sigma}_{ij}^2}}. \quad (11)$$

步驟5. 最後, 我們將進一步探討各校系新年度預測值 \hat{Z}_{j+1} 值的準確度, 以作為落點預測模型 (11) 準確度的衡量準則。

根據迴歸分析理論, 在模型 (2) 下, 給定特定觀測值 z_1^* 後, Z_2^* 的 $(1 - \alpha)100\%$ 預測區間為

$$\hat{Z}_2^* \pm t_{\alpha/2}(K - 2)\sqrt{MSE} \sqrt{1 + \frac{1}{K} + \frac{(z_1^* - \bar{z}_1)^2}{\sum_{k=1}^K (z_{1k} - \bar{z}_1)^2}}.$$

例如九十六年放榜後, 若 (Z_{95}, Z_{96}) 合乎線性模型, 今有某特定值 z_{95}^* , 則 Z_{96}^* 的預測區間為

$$\hat{\beta}_0 + \hat{\beta}_1 z_{95}^* \pm t_{0.025}(K - 2)\sqrt{MSE_{95,96}} \sqrt{1 + \frac{1}{K} + \frac{(Z_{95} - \bar{Z}_{95})^2}{\sum_{k=1}^K (Z_{95k} - \bar{Z}_{95})^2}}$$

其中 $K = 1434$ 為九十五的年校系的總個數, $MSE_{95,96}$ 為利用 (Z_{95}, Z_{96}) 所建立的迴歸分析中的 MSE (mean square error)。但在放榜前, 當年度各校系的最低錄取加權總分是未知的, 所以 Z_{96} 是未知的, 故以 $\hat{\beta}_0 = 0, \hat{\beta}_1 = 1$ 代入得 Z_{96} 之預測值 $\hat{Z}_{96} = Z_{95}$, 同理 $MSE_{95,96}$ 也無法得知, 故本文利用前兩年的迴歸模型的標準誤來替代, 也就是以 $MSE_{94,95}$ 去替代 $MSE_{95,96}$, 因此我們有 Z_{96} 的 95% 預測區間為

$$Z_{95} \pm t_{0.025}(K - 2)\sqrt{MSE_{94,95}} \sqrt{1 + \frac{1}{K} + \frac{(Z_{95} - \bar{Z}_{95})^2}{\sum_{k=1}^K (Z_{95k} - \bar{Z}_{95})^2}}. \quad (12)$$

在第3節中, 我們將透過資料實證分析, 顯示 $MSE_{94,95}$ 與 $MSE_{95,96}$ 差異不大, 此即說明放榜前, 以前兩年的迴歸模型的標準誤來替代的合理性。在聯分會公佈九十六年各校系的最低錄取加權總分後 (放榜後), 我們可以得到 Z_{96} 的真實值, 並依據 Z_{96} 落入預測區間 (12) 的比例, 作為我們提出的落點預測模型準確度的衡量準則。

3 資料分析

在本節中我們將收集九十三至九十六學年度, 聯分會及大考中心所公佈的各校系之指考科目加權權重、各單一考科考生成績累計表, 及各校系最低錄取加權總分

等資料，依循2.5節中的步驟，進行實證分析。首先應用大考中心在放榜前公佈的各考科成績之五標，估計各考科成績分佈的平均數與變異數，進而訂立各校系的評價指標，依此去驗證落點分析的前提假設成立，再建立落點預測模型，最後評估我們提出的落點分析模型的準確度。我們將使用兩年為一組的資料進行分析，如：九十三與九十四年、九十四與九十五年、九十五與九十六年。爲了資料完整性及一致性，我們先剔除各學年度新增校系及加考術科的校系，再將這些校系分成19個學群，對所有校系或依各個學群分類，進行資料分析。

3.1 估計各考科成績分佈的平均數與標準差

[1]中我們直接假設考科成績服從常態分佈且變異數相等，在2.3節我們提出如何應用各單一考科考生成績累計表中的 n 個 p -百分位數(p th percentile)，來評估常態分佈假設是否合理，並進一步估計各考科成績的平均數與變異數。經過多年資料收集與分析，表1詳列各年度之估計值與真值，其中真值來自大考中心網頁[3]上的年度報告研究。表1中可看出平均數的估計值與真實值很接近，在標準差部分，由於每年各考科難易度不同，所以估計值與真實值的差距也有所不同，例如：數乙在94年標準差的估計值就比95年來的精確。經過比較可知，2.3.2節估計方法非常簡單且有效。

3.2 驗證前提假設並建立預測模型

針對各年度 j ，進一步將(6)代入(1)中得到各校系的評價指標 Z_j 如(9)。在建立落點預測模型前，我們應用2.2節中所提出的方法，利用連續兩年的資料 $(Z_{jk}, Z_{(j+1)k})$ ， $k = 1, \dots, K$ 驗證「前後兩年學生對校系的基礎評價不變」的前提假設是否成立。我們針對九十三年至九十六年度的資料進行分析，表2詳列各學群中資料的相關係數。在各學群中，相關係數值大致皆大於0.95，此結果進一步說明前後兩年的評價指標 (Z_j, Z_{j+1}) ，明顯地高線性相關，因此驗證落點分析的前提假設成立。

在假設所有考科成績均具有常態分佈下，接著我們探討2.4節的理論，驗證並建立前後兩年度評價指標的線性迴歸模型。首先根據連續兩年 $(Z_{jk}, Z_{(j+1)k})$ ， $k = 1, \dots, K$ 的散佈圖均顯示模型(2)合理。接著，我們進行模型(2)中虛無假設 $H_0 : \beta_0 = 0$ 與 $H_0 : \beta_1 = 1$ 的統計檢定。令 t_1 表示 $H_0 : \beta_0 = 0$ 的檢定統計量值， t_2 表示

$H_0 : \beta_1 = 1$ 的檢定統計量值, 表3顯示在考量不同學群下, 九十五與九十六年資料 (Z_{95}, Z_{96}) 在模型 (2) 下的相關統計量, 其中 $\hat{\beta}_0$ 均相當靠近 0 且 $\hat{\beta}_1$ 均相當靠近 1, 其餘年度亦有相似結果, 請參看技術報告 [2]。因此我們在各年度放榜前, 以 $\beta_0 = 0, \beta_1 = 1$ 去建立預測模型, 針對各校系 k 採用 (11) 來預測新年度各校系的最低加權錄取總分。

3.3 預測區間及其準確度

本節僅呈現九十五與九十六年度資料 (Z_{95}, Z_{96}) 的處理結果, 其餘年度亦有相似結果, 請參看技術報告 [2]。經由 2.5 節的探討, 我們知道 $MSE_{95,96}$ 於放榜前無法得知, 故以 $MSE_{94,95}$ 去估計 $MSE_{95,96}$, 但放榜後可計算得到 $MSE_{95,96}$, 實際資料分析顯示 $\sqrt{MSE_{93,94}} = 0.1975, \sqrt{MSE_{94,95}} = 0.1680, \sqrt{MSE_{95,96}} = 0.2580$ 。接著, 在取得 Z_{95} 資料後, 即可利用 (12) 式去計算 Z_{96} 的 95% 預測區間。

例如: 國立台灣大學數學系的 $Z_{95} = 3.5602$, 且經由 (Z_{94}, Z_{95}) 資料我們得知, $\sqrt{MSE_{94,95}} = 0.168$, 全部校系個數 $K = 1434, \bar{Z}_{95} = 0.61225, \sum_{i=1}^{1434} (Z_{95i} - \bar{Z}_{95})^2 = 3217.931$, 則該校系 Z_{96} 的 95% 預測區間為

$$\begin{aligned} & 3.5602 \pm 1.962 \times 0.168 \times \sqrt{1 + \frac{1}{1434} + \frac{(3.5602 - 0.61225)^2}{3217.931}} \\ & = (3.230024, 3.890376) \end{aligned}$$

但在聯分會公佈九十六學年度各校系的最低錄取總分後, 即可以得到 Z_{96} 的實際值, 圖 1 為 Z_{96} 與其 95% 預測區間的關係圖。我們觀察到多數落於預測區間的校系, 其評價指標介於 -1.5 ~ 4 之間, 此結果說明了對於評價指標中或中上的校系, 我們的預測值是正確的; 但是針對評價極優或極差的校系, 我們的預測較不準確。表 4 列出九十六學年度所有校系依 Z_{96} 排序後, 前 20 個校系的預測值 \hat{Z}_{96} 、預測區間的上下界、是否有落入預測區間內、最低錄取加權總分的估計值 \hat{T} 與實際值 T 的整理結果, 以供比較, 可更強化對本研究的方法及結果的瞭解。實際上, 在 1434 個校系中共有 1088 個校系落於此預測區間內, 準確比例為 75.87%, 表 5 進一步提供針對部份學校而言, 本預測模型的準確比例, 其餘學校的預測準確率詳列於技術報告 [2]。

4 結論

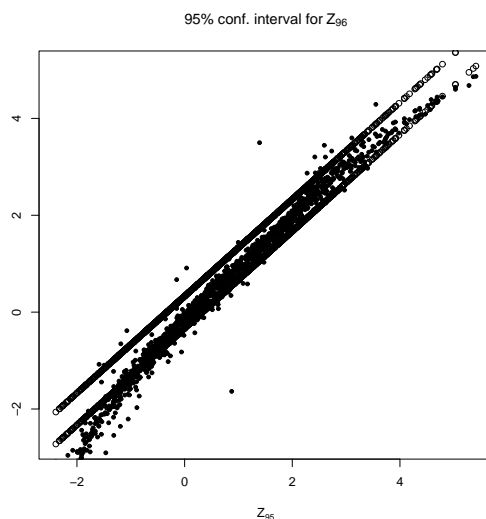


圖 1 Z_{96} 與其 95% 預測區間的關係圖

在大學放榜前，選填志願的考生往往求助於落點分析，而落點分析主要建構在「學生對校系前後兩年基礎評價不變」的前提假設下。自民國九十一年首次實施所謂的「考招分離，多元入學」方案，入學方式逐年修改，我們 [1] 首次應用統計理論，定義各年度各校系的評價指標，並依據該指標驗證落點分析的前提假設成立。進一步應用大考中心與聯分會發佈的各單一考科考生成績累計表，各校系之指考科目與加權方法，及前一年度校系的最低錄取加權總分，寫成一個有效的電腦落點分析軟體，置放於「統計網路學習館」網頁上 [4] 供免費下載。

本文利用統計理論驗證 [1] 中，考科成績具有常態分佈的假設，並應用大考中心所公佈的五標，去估算不同年度各考科成績分佈的平均數與變異數，藉此修正之前的評價指標與預測模型。最後我們引用預測區間來評估本預測模型的準確度。多數校系在新一年度最低加權錄取總分的預測表現還不錯，但對於校系評價較高或較低的校系預測較不精確。尤其在九十六年度，評價指標較低的校系，其預測值與實際值差異相當大。本文提供一個完整的大學指考落點預測理論基礎，經過連續四年實際收集資料後，進行逐步的實證分析，說明我們提出的落點預測模型具有相當高的準確度，並首度完成一個多年大學指考分發結果的具體研究。

參考文獻

1. 任眉眉、陳日昇、詹嘉豪 (2005)。統計與落點分析：大學指考選填志願的輔助利器。中國統計學報, 第43卷第2期, 2005年6月, 165-181。
2. 任眉眉、陳日昇、林家立 (2008)。變異數調整後的大學指考落點預測分析。Technical Report No. 75, Department of Statistics, National Cheng-Kung University。
3. 大考中心的網頁 (<http://www.ceec.edu.tw/>)
4. 成大統計系「統計網路學習館」網頁 (<http://estat.ncku.edu.tw/>)
5. T. W. Anderson (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed., John Wiley & Sons, New York.
6. John Neter, Michael H. Kutner and William Wasserman (1996). *Applied Linear Regression Models*, 3rd ed., McGraw-Hill College.

表 1 93-96 各年度各考科平均數與標準差的估計 $\hat{\mu}$ 與 $\hat{\sigma}$

		國文	英文	數甲	數乙	歷史	地理	物理	化學	生物
96年	$\hat{\mu}$	54.20	30.40	35.00	43.80	53.20	39.40	32.60	43.00	57.00
	$\hat{\sigma}$	14.37	23.03	21.65	23.67	20.19	14.88	27.33	25.68	23.21
	m_e	56	26	33	43	55	40	27	41	56
	μ	54.46	31.12	36.09	44.45	53.40	39.65	32.81	43.55	56.94
	σ	14.47	21.53	21.02	21.92	19.14	14.8	25.10	23.14	20.99
95年	$\hat{\mu}$	51.60	33.20	35.80	54.60	38.80	40.20	26.60	42.40	45.40
	$\hat{\sigma}$	13.55	26.19	21.52	30.54	15.20	17.03	20.32	23.85	21.20
	m_e	52	28	35	56	40	40	22	41	44
	μ	51.61	33.95	36.92	54.55	39.24	40.47	27.09	42.97	45.47
	σ	13.64	23.71	20.66	27.01	14.61	16.16	20.03	22.19	19.75
94年	$\hat{\mu}$	43.60	36.40	33.60	29.20	34.80	36.20	27.80	38.40	45.40
	$\hat{\sigma}$	14.05	26.70	20.51	24.86	18.54	15.88	21.65	29.85	20.83
	m_e	44	34	32	25	35	36	23	34	44
	μ	44.02	37.03	34.51	30.04	35.32	36.57	28.31	38.88	45.93
	σ	13.9	23.79	19.56	23.3	17.42	15.55	20.97	26.64	19.86
93年	$\hat{\mu}$	56.80	30.60	34.80	35.60	30.20	41.00	40.00	33.80	56.80
	$\hat{\sigma}$	14.56	21.01	23.80	22.66	15.89	16.53	27.51	25.50	20.19
	m_e	58	27	30	32	30	42	35	30	57
	μ	56.50	30.50	34.50	36.00	30.50	41.00	39.50	34.00	56.00
	σ	14.41	20.26	22.83	22.06	15.14	16.03	24.95	23.28	19.03

¹ 當年度各考科之平均數 μ 、標準差 σ 之資料, 為大考中心網頁所提供 [3]。

² $\hat{\mu}$ 、 $\hat{\sigma}$ 為利用本文方法所估計出的當年度各考科之平均數與標準差。

³ $m_e = x_{p_3}$ 為中位數。

表 2 九十三年至九十六年度評價指標的相關係數

分類	K(總數)	$r_{Z_{93},Z_{94}}$	K	$r_{Z_{94},Z_{95}}$	K	$r_{Z_{95},Z_{96}}$
所有校系	1373	0.9909	1400	0.9937	1434	0.9883
大眾傳播學群	41	0.9891	44	0.9975	46	0.9928
工程學群	200	0.9957	199	0.9937	207	0.9934
文史哲學學群	73	0.9929	76	0.9933	77	0.9927
外語學群	100	0.9874	102	0.9960	108	0.9879
生命科學學群	45	0.9960	45	0.9977	47	0.9932
地球環境學群	28	0.9884	27	0.9976	8	0.9963
法政學群	69	0.9948	72	0.9955	71	0.9951
社會心理學群	73	0.9889	77	0.9977	81	0.9741
建築設計學群	40	0.9291	36	0.9870	36	0.9664
財經學群	123	0.9937	125	0.9928	121	0.9909
教育學群	55	0.9921	53	0.9950	54	0.9881
資訊學群	121	0.9916	125	0.9925	129	0.9857
農林漁牧學群	36	0.9917	38	0.9905	38	0.9826
管理學群	91	0.9867	105	0.9926	107	0.9866
數理化學群	123	0.9896	122	0.9772	126	0.9931
醫藥衛生學群	134	0.9962	135	0.9970	138	0.9946
藝術學群	11	0.9921	10	0.9792	11	0.9940
體育學群	6	0.8857	5	0.9642	5	0.9792
其他學群	4	0.9999	4	0.9998	4	0.9966

表 3 九十五與九十六年在模型 (2) 下的相關統計量

分類	總數	$\hat{\beta}_0$	$\hat{\beta}_1$	t_1	t_2	R^2
所有校系	1434	-0.2677	1.1153	-36.400	25.3400	0.9770
大眾傳播學群	46	-0.1846	1.1685	-6.4549	7.9213	0.9856
工程學群	207	-0.3411	1.1281	-21.9484	14.1143	0.9869
文史哲學群	77	-0.1810	1.0975	-8.3685	6.3465	0.9855
外語學群	108	-0.2210	1.1414	-8.2182	8.1179	0.9759
生命科學學群	47	-0.4249	1.0957	-13.0732	4.9965	0.9864
地球與環境學群	28	-0.3141	1.1299	-9.3383	6.7609	0.9925
法政學群	71	-0.2090	1.1426	-9.1373	11.3500	0.9917
社會與心理學群	81	-0.2668	1.1300	-6.8789	4.6379	0.9536
建築與設計學群	36	-0.1285	1.0892	-2.4437	1.7942	0.9339
財經學群	121	-0.1794	1.1873	-9.2423	13.4613	0.9839
教育學群	54	-0.2009	1.1522	-6.6060	6.1157	0.9763
資訊學群	129	-0.3385	1.1907	-12.6481	10.5580	0.9716
農林漁牧學群	38	-0.3255	1.1079	-6.3032	3.0902	0.9655
管理學群	107	-0.1885	1.2197	-7.1928	11.1511	0.9733
數理化學群	126	-0.3128	1.0870	-17.4291	7.5277	0.9862
醫藥衛生學群	138	-0.2585	1.0197	-11.9322	2.1497	0.9892
藝術學群	11	-0.2277	1.1229	-3.3474	2.9788	0.9880
體育學群	5	-0.3420	1.0942	-2.9535	0.7197	0.9588
其他學群	4	-0.0237	0.9906	-0.1841	-0.1635	0.9933

表 4 依 Z_{96} 排序, 前 20 個校系的相關資料

學校	科系	Z_{96}	\hat{Z}_{96}	預測區間上界	預測區間下界	是否有落入	\hat{T}_{96}	T_{96}
國立臺灣大學	醫學系 (自費)	4.8680	5.4109	5.7418	5.0801	0	555.93	525.45
國立臺灣大學	醫學系 (公費)	4.8605	5.3621	5.6929	5.0312	0	553.18	525.03
國立陽明大學	醫學系 (自費)	4.6789	5.2818	5.6126	4.9510	0	548.68	514.84
國立成功大學	醫學系	4.6426	5.0311	5.3617	4.7004	0	534.60	512.8
國立陽明大學	醫學系 (公費)	4.6006	5.0311	5.3617	4.7004	0	534.60	510.44
長庚大學	醫學系	4.4458	4.7895	5.1201	4.4590	0	592.34	570.27
國立臺灣大學	牙醫學系	4.4406	4.6902	5.0207	4.3596	1	515.47	501.46
國立臺灣大學	電機工程學系	4.3891	4.5802	4.9106	4.2497	1	429.29	419.52
臺北醫學大學	醫學系 (自費)	4.3615	4.6845	5.0150	4.3539	1	515.15	497.02
臺北醫學大學	醫學系 (公費)	4.3326	4.6784	5.0089	4.3478	0	514.81	495.4
慈濟大學	醫學系 (公費)	4.3007	4.5935	4.9240	4.2630	1	510.04	493.61
國立臺灣師範大學	特殊教育學系 (公費)	4.2903	3.5521	3.8822	3.2220	0	413.99	449.7
高雄醫學大學	醫學系 (公費)	4.2662	4.5144	4.8448	4.1839	1	505.60	491.67
國立臺灣大學	物理學系	4.2090	4.3665	4.6969	4.0361	1	418.37	410.32
國立陽明大學	牙醫學系	4.1839	4.3683	4.6987	4.0379	1	541.63	530.33
高雄醫學大學	醫學系 (自費)	4.1812	4.5627	4.8932	4.2322	0	508.32	486.9
國立臺灣大學	材料科學與工程學系	4.1625	4.2891	4.6195	3.9588	1	414.41	407.94
國立交通大學	電機資訊學士班	4.1511	4.2756	4.6060	3.9453	1	413.72	407.36
慈濟大學	醫學系 (自費)	4.0861	4.4691	4.7995	4.1387	0	503.06	481.56
輔仁大學	醫學系	4.0702	4.3883	4.7187	4.0579	1	498.52	480.67

表 5 九十六學年度部分大學 Z_{96} 落在預測區間內的百分比例

學校名稱	落入預測區間個數	校系總數	比例 (%)
國立臺灣大學	55	63	87.30
國立臺灣師範大學	20	22	90.91
國立中興大學	32	32	100.00
國立成功大學	37	38	97.37
東吳大學	14	21	66.67
國立政治大學	38	42	90.48
高雄醫學大學	19	20	95.00
中原大學	29	30	96.67
東海大學	28	36	77.78
國立清華大學	20	21	95.24
中國醫藥大學	19	20	95.00
國立交通大學	22	24	91.67
淡江大學	44	44	100.00
逢甲大學	24	32	75.00
國立中央大學	23	23	100.00
中國文化大學	43	54	79.63
靜宜大學	20	24	83.33
大同大學	6	13	46.15
輔仁大學	43	49	87.76
國立臺灣海洋大學	16	17	94.12
國立高雄師範大學	14	14	100.00
國立彰化師範大學	16	16	100.00
國立陽明大學	6	8	75.00

Statistical Forecasting After the College Entrance Examination with Real Data

Mei-Mei Zen¹, Zue-Seng Chen² and Chia-Li Lin¹

¹DEPARTMENT OF STATISTICS, NATIONAL CHENG-KUNG UNIVERSITY

²Computer and Network Center, National Cheng-Kung University

ABSTRACT

After the yearly College Entrance Examination, an examinee always faces the problem of selection and priority of certain departments for advance study. Under the assumption that the test scores of all nine subjects follow normal distributions with the same variance, we [1] established an index for each student to represent the possibility of entering any interested department. In this paper, we make the statistical inference about the normal assumption first. Then we derive the parameter estimators of the normal distribution. Using the estimates, we adjust our previous index to be more reasonable. Moreover, our forecasting model is supported through multivariate analysis and prediction intervals. As a pioneer study, a four-year real data is collected and applied to support our theoretical results.

Key words and phrases : Weighted score, normal distribution, multivariate analysis, linear regression, prediction interval, accuracy.

AMS 2000 subject classifications: Primary 62J05; secondary 62P25.