第一章

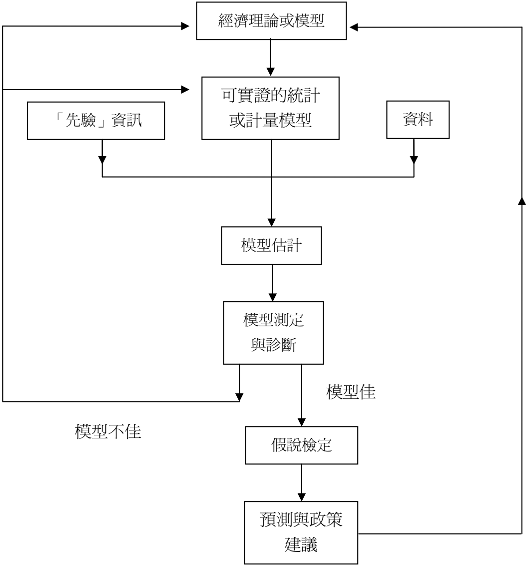
緒論

September 29, 2010

1.1 簡介: 什麼是計量經濟學

- 作為經濟學門的一個分支, 計量經濟學 (Econometrics), 基本上是探討將概念性的經濟理論藉由實際的資料予以數量化的研究學門。
- 字面上的意義是指「經濟的衡量 (economic measurement)」, 較精確的說, 則計量經濟學是探討「實證研究方法」的研究學門。
- 計量經濟學強調經濟關係的「數量化」, 所以其目的在於經濟「理論」與「衡量」的結合。

- 純粹概念性的理論顯然在實際問題的解釋與應用上是有限的; 而缺乏理論基礎的統計分析則無法提供一個「因果關係」的說明, 同樣, 其說服力也是有限的, 或甚至可能是錯誤的。


經濟理論或模型
可實證的統計
或計量模型
「先驗」資訊
資料
模型估計
模型測定
與診斷
模型佳
模型不佳
假設檢定
預測與政策
建議

- 計量經濟學包括四個主要成分: 理論、資料、計量方法, 與計算工具(即執行的軟體與硬體)。
- 基本上, 一個實證研究應從一個概念性的經濟理論出發。例如, 經濟學裡的"需求函數"認為需求的數量Q是價格P的遞減函數:

$$
\begin{aligned}
Q &= g(p), \\
\frac{\partial Q}{\partial P} &= \frac{\partial g(\cdot)}{\partial P} < 0
\end{aligned}
$$

但是我們所知的也僅限於此, 至於Q與P之間的函數型式$g(\cdot)$為何, 我們卻一無所知。

- 要將「概念的需求函數」予以「具體化」, 第二步就是需要假設一個可實證的統計或計量模型。計量模型的決定就是要決定一個明確的函數型式, 這可能有需配合上研究者的一些 "先驗" 資訊 (或是經驗), 以及可蒐集到的相關資料。
- 例如, 我們認為需求函數有如下的關係:

$$Q_i = \beta_1 + \beta_2 P_i + \beta_3 I_i + \varepsilon_i$$

上式中,Q_i、P_i、I_i分別代表某個觀察值i對應的需求數量、價格, 與所得水準。

- 「模型估計」的下一步驟就是要對模型估計的結果作進一步測定與診斷。最主要的就是看模型估計的結果是否與理論的預測一致。如果不佳的話, 則可能有必要對理論作修正, 或是對計量模型作修正。
- 需要注意的一點是, 在將概念性的理論轉換為實證的模型時, 常常會需要加一些額外的假設, 而這些假設則未必是原有理論模型的假設。因此, 如果實證結果不支持理論, 則有可能是理論錯誤, 也有可能是我們額外所加的假設是不恰當的。

- 假若模型的配適佳的話, 則我們可進一步作假說檢定, 以瞭解理論是否得到資料的支持。
- 下一步, 我們就可進一步作預測與政策建議, 也可進一步對經濟理論或模型作回饋與修正。如此, 就完成了計量經濟 (或實證研究) 的一個循環 (cycle)。
- 綜言之,「計量經濟學」係藉由統計工具將概念性的經濟理論付諸實際的一項學科。

- 要注意的是, 統計只能告訴我們「相關」, 而不能講「因果」。
- 從邏輯的觀點來看, 一個命題「$p \Rightarrow q$」的對偶命題是「$\sim q \Rightarrow \sim p$」
- 實證作法上, 我們一般是先有一個理論或假說 (即p), 然後進一步認為現實的「資料」或現象 (即q) 應該會與我們的理論一致。假若實證的結果果然與我們的理論預期的相同 (即q成立), 則我們就認定理論成立。嚴格而言, 這在邏輯上是有問題的。
- 倘若結果不支持理論的預期, 我們的確可說理論不成立 (在此對偶命題是成立的)。但是我們卻無法確定q的成立是因為p成立的結果; 有可能其他的理論也會導致q的成立。

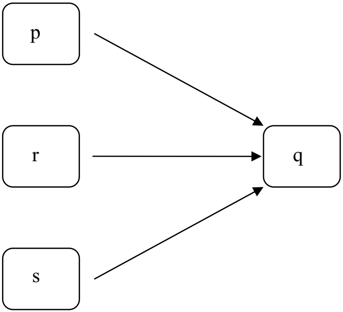


Figure: 1.2: 充分、必然關係圖解

- 圖 1.2 顯示p、r、s都可以導致q的成立, 因此當q成立時, 有可能是p、r、s三者其一或全部成立, 這是我們無法驗證的。不過如果q不成立, 則可知p、r、s三者都不可能成立。
- 所以統計可駁斥 (refute) 一項理論或假說, 但不能「證明」或是 verify。所以至多也只能是研究的輔助工具。

- 總括來說, 我以為「建模 (modeling)」, 亦即建立或尋找適當的模型來檢驗理論或假說, 是最重要的一項工作。
- 正確的計量方法, 包括適當的估計與檢定方法, 與統計概念也是很重要的, 只是相對之下, 有時技術性的細節 (technical details) 或許就顯得不是那麼地重要了。

1.2 一個簡單例子

- 假如我們想分析台積電股票報酬的決定因子, 而根據財務理論, 我們知道台積電股票的報酬與市場大盤有某個一定的關係。
- 令r_t代表台積電股票在某t日的報酬, 而r_{mt}代表加權股價指數在t日的報酬。
- 假如在某日, 大盤漲了1%, 而台積電漲了3%, 如圖1.3中的A點。而第二天, 二者的報酬為(r_t, r_{mt})=(1%,2%), 如圖中的 B 點。
- 假如我們相信二者有線性關係, 亦即

$$r_t = a + br_{mt},$$

可求出a=5%,b=-2%。

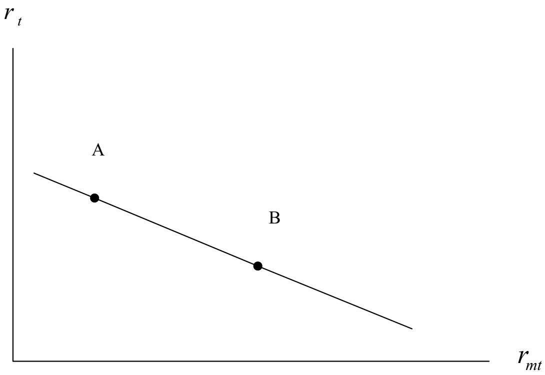


Figure: 1.3: 二個點的關係

- 這個結果告訴我們, 假如大盤漲 5%, 則台積電將下跌$r_t = 5\% - 2 \times 5\% = -5\%$。這顯然與我們的直覺不符。
- 假如第三點是(r_t, r_{mt})=(0.5%,0.5%), 如圖 1.4 中的C點。
- 事實上, 只要再多加一個 "參數" 就仍然可以完全捕捉A,B,C三點。最簡單的方式就是用一個一元二次方程式來配適 (fit) 此一關係:

$$r_t = a + br_{mt} + cr_{mt}^2$$

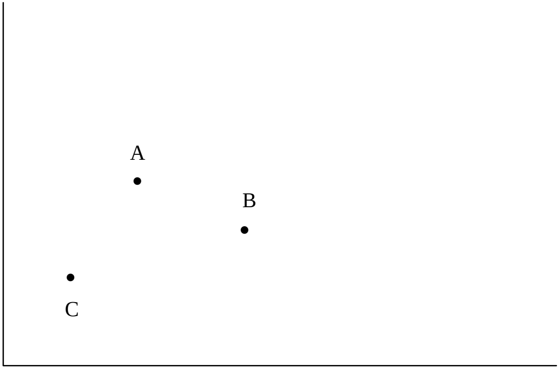


Figure: 1.4: 三個點的關係

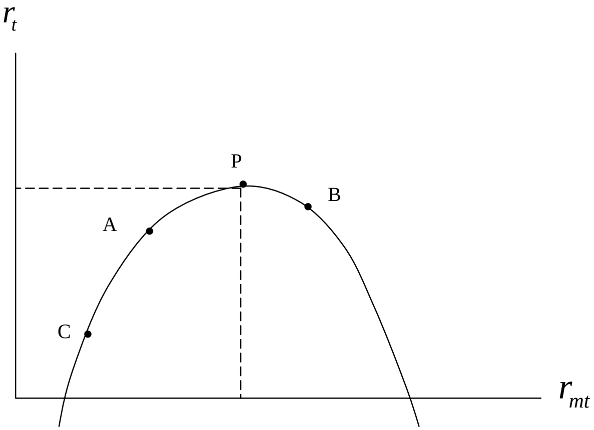


Figure: 1.5: 二次式的配適結果

- 重新配適的結果如圖 1.5 所示。新的二次式完全地將 A,B,C 三點 “包含” 在內。但是二次式的配適結果卻隱含大盤大漲時，台積電報酬卻是下跌的，而且二次式也表示報酬有個上限 (圖5之 P 點)，二者都與我們的直覺不同。

- 假設我們又取得一D點, 其實我們還是可以再增加一個參數來求得完全的配適 (exact fitting):

$$r_t = a + br_{mt} + cr_{mt}^2 + dr_{mt}^3,$$

- 但是我們相信上述的配適結果嗎? 我想是值得存疑的。其實只要模型中的參數夠多, 我們總是可以完全地 fit 我們手邊的資料。但如同上面這些圖顯示的, 如此的配適常常得到很奇怪, 而且在預測上又很不準確的結果。

- 事實上, 從統計的角度來看, 每一筆資料都攜帶著「真正模型」的相關資訊 ,而每一筆資料也代表著一個「自由度」。
- 由於我們未必知道真正的關係為何 (可能永遠也不能知道), 因此隨便假設一個模型來 "規範" 或限制變數與變數間的關係, 反而會是一種錯誤。
- 模型中每加一個參數就相當於對變數間的關係多加一個限制, 因此也使得手邊樣本的自由度少掉一個。

1.2.1 淺談「自由度」

- 「自由度」代表的是一個「點」可以“自由” 移動的象度或維度 (dimension)。
- 例如，在一條水平線上的點只能向左或向右移動，因此我們只要給定一個值就可以決定其位置，如$x = 3$。同理，在 XY 平面上的一個點則需要其X軸與Y軸的座標值才能確認其位置，所以其自由度為 2。
- 但假若我們加上$2x + y = 5$的限制呢？我們知道一個點只能在$2x + y = 5$的直線上移動，因此自由度只剩下一個。換句話說，每加一個限制，自由度就會少掉一個。

- 舉例來說，若(x, y)代表甲與乙二人口袋裡的錢，則我們知道在未有任何資訊下，(x, y)的自由度為2。但若我們進一步得知二人的總金額為10元 (這就是一個限制)，則我們只要知道其中任一個人的金額，就能得知整個“樣本”(在這裡就是二人) 的情況。
- 假如我們進一步知道甲比乙多2元呢？則可知自由度還要再減少一個，這時我們有二個限制式：

$$\begin{aligned} x + y &= 10 \\ x - y &= 2, \end{aligned}$$

可知$(x, y) = (6, 4)$，這時(x, y)被限定在一個點上，自由度為0。

- 依此類推，可知三度空間中的點有三個自由度，也就是我們必須知道其(x, y, z)座標才能確認其位置。換句話說，一個n度空間中的點基本上有n個自由度。要確切知道其位置，就必須知道其n個座標所對應的值才行。

1.2.2 計量經濟的作法 (An Econometric Treatment)

- 假設實際上我們蒐集了T筆報酬的資料,$\{(r_1, r_{m1}), \cdots, (r_T, r_{mT})\}$。令

$$y = \begin{pmatrix} r_1 \\ \vdots \\ r_T \end{pmatrix}, x = \begin{pmatrix} r_{m1} \\ \vdots \\ r_{mT} \end{pmatrix}$$

  則可知y是 “T度” 空間中的一個點 (也可視為一個向量), 自由度有T個。

- 我們將(r_t, r_{mt})繪於圖 1.7。

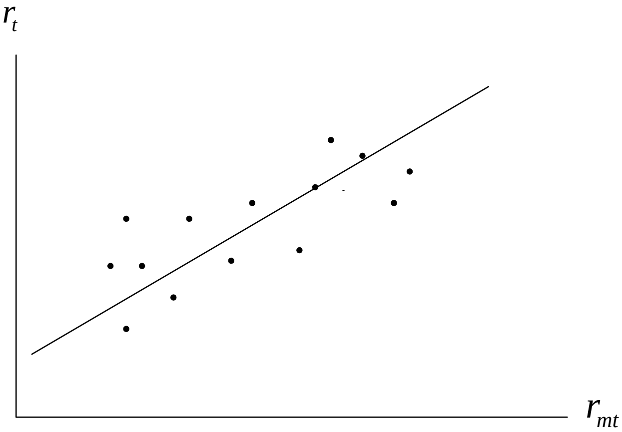


Figure: 1.7: (r_t, r_{mt})的分佈圖

- 如同我們前面所討論的, 要完全配適r_t與r_{mt}的關係, 只要採用一個「一元$(T-1)$次方程式」就行:

$$r_t = a_0 + a_1 r_{mt} + a_2 r_{mt}^2 + \cdots + a_{T-1} r_{mt}^{T-1}$$

但是這樣的作法並不恰當, 而這樣的假設背後隱含r_t是完全由r_{mt}單一因子所決定。這當然不是一個合理的假設。

- 影響股票報酬有許多, 而事實上我們又不可能將所有的因素包含進來(因為蒐集資料需要成本, 而資源或預算卻是有限的), 因此也只能儘量地將重要的相關變數包含進來。如圖 1.7 所顯示的, 我們可看出r_t與r_{mt}間仍有一個共同趨勢, 而此一趨勢則又是線性的。
- 由於變數間的關係並非完全線性, 最好的方式就是引入一個誤差項, 或是干擾項, 來 "吸收" 所有非模型內變數的影響。這個關係可表達為:
$$r_t = \alpha + \beta r_{mt} + \varepsilon_t, t = 1, \cdots, T. \qquad (1)$$
ε_t即所謂的誤差項, 代表r_{mt} 以外的因素 (如天氣、國外股市等$\cdots$) 對r_t的影響, 因此,ε_t基本上是與r_{mt}不相關的, 也就是$cov(\varepsilon_t, r_{mt}) = 0$。$\varepsilon_t$是一個隨機變數, 而不是一個固定值。

- 這個模型就是最簡單的「簡單線性迴歸模型 (simple linear regression model)」。模型的特點就是只引進一個解釋, 因此需要估計的參數只有二個 (α與β)。
- 誤差項ε_t的引入是統計學的特色。最簡單的作法是假設ε_t服從一個獨立同態 (independently and identically distributed, iid) 的分配, 而且其平均值為零, 而變異數為一個固定值 (令其變異數為σ_ε^2)。也就是上式(1.1) 可拆解為二部分:

$$r_t = (\alpha + \varepsilon_t) + \beta r_{mt}$$

- βr_{mt}代表r_t與r_{mt}有關的部分, 也就是引進解釋變數r_{mt}於模型中後, r_t的報酬 "被解釋掉" 的部份。第一項$(\alpha + \varepsilon_t)$則代表r_t仍未被解釋的部份。所以基本上,α(稱為「截距項」) 係捕捉其他 "非模型變數" (也就是未被納入模型中的變數) 的影響的平均值。

- 將上式兩邊取變異數，得

$$
\begin{aligned}
var(r_t) &= var(\alpha + \varepsilon_t) + var(\beta r_{mt}) + 2\ cov(\varepsilon_t, \beta r_{mt}) \\
&= var(\varepsilon_t) + \beta^2 var(r_{mt})
\end{aligned}
$$

- 第一個等式中$cov(\varepsilon_t, \beta r_{mt}) = 0$，因為我們假設誤差項$\varepsilon$與解釋變數$r_{mt}$是無相關的。
- 上式中,$var(r_t)$是原本變數r_t的總變異。在引進解釋變數r_{mt}後，其變異與r_{mt}有關的部份為$\beta^2 var(r_{mt})$，而$var(\varepsilon_t)$則為仍無法被解釋的變異。

- 如同我們前面說過的, 每加一個參數都代表多一個限制, 會喪失一個自由度。喪失自由度的結果是使得估計與預測的準確度受到影響。但是加入參數同時也代表對變數變動的 "結構" 作了假設或限制。假若此一假設是正確的, 則因爲此一 "結構" 假設所帶來的估計或預測利益將可彌補喪失自由度所帶來的成本。
- 綜言之, 計量經濟學或統計學的作法或重要原則就是「精簡」。精簡原則就是尋找一個既精準(precise) 而又簡約 (parsimonious) 的模型, 以對變數間的關係作最好的推估與分析。

1.3 資料型態與模型分類

- 實證上資料類型可分為三類, 分別是時間序列 (time series)、橫斷面 (cross section), 與追蹤類型 (panel) 三種。
- 時間序列的資料是樣本的觀察值間是以時間點的不同來作區隔的。例如, 一段期間 (如 1990-2000 年) 的大盤指數日資料。
- 若資料不是以時間點來作區隔, 則可稱之為橫斷面資料。一般橫斷面資料比較是指一固定時點的不同觀察值。例如, 上個月不同縣市的失業率。

- 追蹤資料則同時包含了二種資料特性。例如, 過去一年每個縣市的每月失業率就同時包含了時間與橫斷面的特性。不過, 一般而言, 追蹤資料是指“大”的橫斷面與“短”的時間序列。
- “小”的橫斷面與“長”的時間序列的資料型態, 則一般只是稱為合併型態資料 (pooled data)。在分析上, 主要以所謂「系統模型」來處理。
- 而追蹤資料則會以所謂的「追蹤資料模型 (panel data models)」來分析。

就不同屬性進行分類

- 就解釋變數 (independent variables) 的個數，來分：
  - 簡單迴歸模型 (simple regression model)：即僅一個解釋變數。如下例

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$$

  - 複迴歸模型 (multiple regression)：亦即解釋變數數目超過一個以上。

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t.$$

- 就是否線性模型來分:
  - 一般線性模型 (general linear models): 上一頁的二個模型就是線性模型，但是下面這個模型是線性的嗎?

    $$y_t = \alpha + \beta x_t^2 + \varepsilon_t$$

    答案是:Yes! 因為在這裡，我們可以設立一個新變數，如$z_t = x_t^2$，並另模型如下:

    $$y_t = \alpha + \beta z_t + \varepsilon_t$$

    則顯然仍是線性模型。
  - 非線性模型 (non-linear model): 當解釋變數無法經由轉換成為線性時，即為非線性模型。如下例:

    $$y_t = \alpha + \beta x_t^\gamma + \varepsilon_t$$

    非線性模型的估計與分析較線性模型複雜。基本上，仍是透過適當的「線性化 (linearize)」來處理。

- 就「方程式」的數目來分，可大致分為以下幾類：
  - 單一方程式 (univariate equation) 模型：

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

  - 多方程式組 (sets of models)：又大致可分二類，
    1. 聯立方程式模型 (simultaneous models)：聯立方程式模型中，一方程式的被解釋變數 (即所謂的內生變數 (endogenous variable)) 可能成為其他方程式之解釋變數。例如：

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_1 + \beta_2 z_t + e_t \\ z_t &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \eta_t \end{aligned}$$

       z_t同時是解釋變數，也是被解釋變數。

2. 彷彿無相關迴歸模型 (seemingly unrelated regression; SUR) model: 各方程式的解釋變數皆不包括其他方程式的內生變數, 亦即這些解釋變數都是外生的 (即所謂的外生變數 (exogenous variable)); 可視為聯立方程式的特例。

$$
\begin{aligned}
y_{1t} &= \alpha_0 + \alpha_1 z_{1t} + \alpha_2 z_{2t} + \cdots + \varepsilon_{1t} \\
y_{2t} &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \varepsilon_{2t} \\
y_{3t} &= \gamma_0 + \gamma_1 x_{1t} + \gamma_2 z_{3t} + \cdots + \varepsilon_{3t} \\
\vdots \quad &\vdots \qquad\qquad \vdots \\
y_{nt} &= \delta_0 + \delta_1 x_{1t} + \delta_2 x_{2t} + \cdots + \varepsilon_{nt}
\end{aligned}
$$

多變量迴歸模型 (multivariate model) 則又是 SUR 模型的一個特例。

3. 追蹤資料模型。

- 就被解釋變數是否為連續性隨機變數而言:
  - 若是連續函數，如常態分配函數，即為傳統的 normal linear regression model。
  - 若是離散的 (discrete)，例如$y = 0$或1 即為二元選擇模型 (binary choice model)，常見的有 probit 與 logit models。
  - 若被解釋變數的範圍受到截斷，或有所限制，如 Tobit model 或 truncated regression model.