시스템槪念과 計量模型設定

呂 運 邦

▷ 目 次 ◁

Ⅰ. 序
Ⅱ. 一般的 槪念
Ⅲ. 計量模型
Ⅳ. 統計「소프트웨어」
Ⅴ. 結言

Ⅰ. 序

우리들의 日常生活이나 國家政策 樹立過程에 있어서 대부분의 경우에 意思決定(decision making) 혹은 選擇(choice)을 할 필요가 있다. 意思決定 자체는 利害의 得失을 내포하고 있기 때문에 올바른 意思決定이 要求되며 이를 위해서는 결정이 행해지는 대상에 대한 정확한 理解 및 최대한의 情報가 뒷받침되어야 한다.

정확히 1미터의 길이를 갖은 物體란 槪念上으로만 存在할 뿐 실제로는 存在하지 않으며 現實上의 한 變數는 다른 많은 變數들의 函數로 생각되어진다는 것은 쉽게 首肯이 가는 일이다. 그러므로 많은 計量的 槪念들은 抽象的으로만 定義될 뿐 실제의 對象物에 대한 計量은 近似値(approximation)로 사용되고 있으며, 計量槪念에서 한걸음 더 나아가 어떠한 시스템의 有機的 관계를 硏究할 경우에도 그 실제의 시스템을 簡略化(simplification)한 模型(model)을 設定하고 이를 토대로 얻어진 결과를 이용하여 意思決定을 하는 경우가 많다. 이와 같이 近似値로 사용되는 資料가 實際値와 얼마나 近接하며 혹은 設定된 模型이 실제의 시스템을 얼마나 잘 나타내고 있는가 하는 問題는 意思決定의 옳고 그름은 물론 利害의 得失과 直結됨은 자명한 일이다.

本稿에서는 우리가 硏究對象으로 택하는 실제의 시스템에 대한 理解方法과 이를 硏究하는 데 있어 필수적인 構成要素의 하나인 模型에 대하여 分類를 좀더 細分하는 한편 그 特性을 알아 봄으로써 社會科學 硏究分野의 模型設定에 도움이 될 수 있는 보다 合理的인 接近方法을 提示하고자 하며 具體的인 模型의

筆者：韓國開發硏究院 硏究委員

「파라미터」推定方法 등은 매우 多樣하고 複雜하므로 생략하였다.

第Ⅱ章은 우리의 分析 및 硏究對象이 되는 시스템과 模型의 定義와 아울러 시스템 시뮬레이션의 각 段階를 설명하였으며 第Ⅲ章에서는 社會科學硏究에서 가장 자주 쓰이는 計量模型에 대해서 이들 模型의 根幹을 이루는 基本假定을 중심으로 分類하여 計量模型들 각각의 特性을 명확히 하려고 힘썼다. 특히 一般線型模型과 線型回歸模型은 흔히 混同하기 쉬운 模型들이므로 그 구분을 强調하였다. 마지막 第Ⅳ章에서는 여러 模型의 分析에 필요한 컴퓨터 「소프트웨어」를 소개함으로써 硏究業務의 遂行에 參考가 되도록 하였다.

Ⅱ. 一般的 槪念

1. 시스템

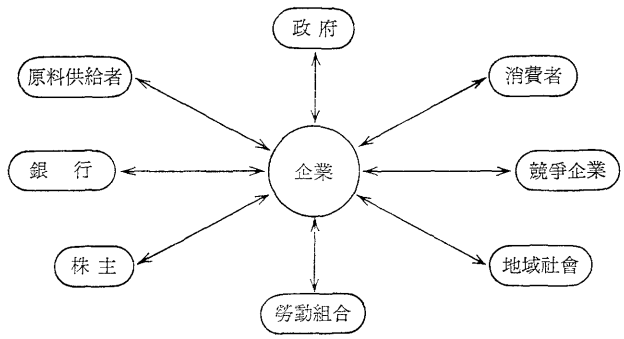
많은 分野에서 시스템(system)이란 用語가 사용되고 있다. 예를 들어 企業에서는 在庫시스템, 分配시스템, 生産시스템 등과 意思決定시스템, 情報시스템 등의 用語가 사용되고 있으며 또한 自然科學, 社會科學 分野에서도 다양한 형태로 사용되고 있다. 이러한 시스템을 조금은 모호하지만 一般的으로 定義하면 다음과 같다.

한 시스템이란 주어진 시스템元素(system element or object)들의 集合(set), 그 元素들의 屬性(attribute or characteristic), 元素들 및 屬性들간의 關係(relation)로 構成된다. 시스템元素들은 機械, 原料, 製品, 雇傭人 등 物質的인 것일 수도 있고 利潤, 生産基準, 價格 등 抽象的일 수도 있다. 시스템이 集合과 다른 점은 단순히 對象物의 모임을 集合이라 하는 데 반해 시스템은 이 集合에 關係나 屬性들이 주어진다는 것이다. 예를 들어 한 硏究所의 職員이 300名일 때 職員全體는 단순한 集合이 되며 個個人에게 어떤 補職, 役割, 關係 등이 주어지면 硏究所시스템이 된다. 한편 시스템의 活動(activity)이란 시스템의 變化를 일으키는 過程을 말하며 시스템狀態(system state)는 한 時點에서의 元素, 屬性, 活動 등을 意味한다.

우리의 硏究對象은 시스템元素들이 주어졌을 때 이들이 어떠한 關係로 作用하는가를 알아내는 것으로서, 예를 들면 販賣量과 廣告費가 밀접한 關係가 있다는 前提下에 이들의 關係를 分析하여 알아내는 것 등이 硏究對象이 될 수 있다. 물론 시스템元素들의 關係가 여러 개 存在하여 모두 識別할 수 있다 하여도 시스템에 영향을 주는 關係만이 分析의 對象이 된다.

한 시스템은 가끔 外部의 變化에 의하여 영향을 받게 되므로 어떤 시스템이 定義되면 硏究의 目的에 따라 시스템環境(system environment)이 정확히 糾明되어야 하는바 이를 定義하면 다음과 같다. 시스템環境이란 시스템에 속하지 않은 元素들의 集合으로서 그 元素들의 屬性變化가 시스템에 영향을 주고 또한 시스템의 變化는 그 元素들의 屬性을 變化시킨다. 시스템과 시스템環境의 구분은 시스템을 分析하는 各者의 편의에 따라 정해지나 經營活動(management activity)에 실질적으로 從屬되느냐 아니냐에 따라 구분하는 것이 一般的이다. 예를 들어 企業시스템에서의 消費

〔圖 1〕 企業시스템과 시스템環境



者는 嗜好와 所得이라는 屬性의 側面에서는 시스템環境에 속한다고 할 수도 있고 企業시스템의 意味를 좀더 擴張시키면 이 시스템에 포함될 수도 있다(圖 1 參照).

시스템과 시스템環境의 定義로부터 한 시스템이 여러 개의 部分시스템(subsystem)으로 分解가 가능함을 알 수 있다. 企業시스템의 경우 主要 部分시스템에서는 生產시스템, 人力시스템, 會計시스템, 分配시스템으로 나누어질 수 있고, 이 뿐만 아니라 經營情報시스템(MIS; Management Information Systems)과 같이 全體部署에 관계되는 部分시스템이 존재할 수도 있다. 또한 한 部分시스템의 元素는 다른 部分시스템環境에 속할 수도 있고 동시에 다른 여러 部分시스템의 元素인 경우도 있으며 더욱 한 部分시스템은 擴張시스템의 元素일 수도 있어서 시스템간의 階層을 이루게 된다. 이러한 構造的 複雜性 때문에 어떻게 시스템과 시스템環境을 定義하느냐 하는 것은 硏究의 目的에 따라 달라지게 된다.

시스템活動은 內生活動(endogenous activity)과 外生活動(exogenous activity)으로 나눌 수 있는데 이 중 內生活動이란 시스템의 內部에서 일어나는 活動을 말하며, 外生活動이란 시스템에 영향을 주는 시스템環境에서의 活動을 意味한다. 한편 시스템은 外生活動의 存在與否에 따라 닫힌시스템(closed system)과 열린시스템(open system)으로 구분한다. 열린시스템은 시스템環境과의 相互作用을 단절시키거나 相互作用이 일어나는 시스템環境의 일부를 시스템에 포함시킴으로써 닫힌시스템으로 만들 수 있다.

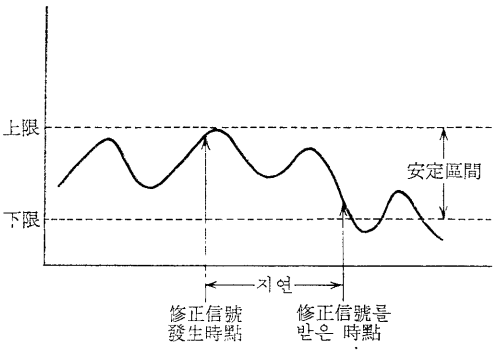
또한 시스템活動의 결과가 시스템의 入力에 의해서 완전히 결정되는 경우 이를 決定的活動(deterministic activity)이라 하고 시스템活動의 영향이 여러 가지 결과의 可能性을 가지고 변할 때는 確率的活動(stochastic activity)이라 부른다. 確率的活動의 任意性은 確率分布(probability distribution)의 형태로 表現되며 이는 시스템環境의 活動을 포함하는 경우가 많다. 시스템活動의 發生自體가 시스템 制御上에 있다면 이 活動은 內生的이며 이것이 任意的이면 시스템環境의 活動이라 보아야 한다. 예를 들면 工場에서 電力으로 機械를 稼動할 때 機械稼動時間은 비록 어떤 確率分布를 갖는지는 알 수 없지만 시스템 制御上에 있으므로 內生活動이며 發電所나 送信所의 고장에 의한 임의의 時間동안의 電力斷切은 外生活動의 結果라고 할 수 있다.

한 時點에서의 시스템狀態는 屬性들의 값을 觀測한 觀測値로 表現되며 이들을 變數로 볼

때 變數의 값은 時間의 흐름에 따라 觀測된다. 이러한 모두 變數들이 常數로 남아 있게 되는 경우나 혹은 어떤 區間을 벗어나지 않는 경우의 시스템을 安定的시스템(stable system)이라 하며 그렇지 않은 경우의 시스템을 不安定的 시스템(unstable system)이라 한다. 不安定的 시스템은 時間이 갈수록 그 振幅이 커지거나 發散하는 정도가 커지는 것이 一般的이다.

外部의 충격이 없을 때 시스템狀態가 변하지 않고 그대로 남아 있으면 이를 均衡狀態(equilibrium state)라 하는데 外部의 충격이 있더라도 均衡狀態로 돌아가면 이 시스템은 安定的인 것이다. 또한 시스템의 作用을 入力(input), 變換過程(transformation process), 出力(output)의 세 단계로 구분하여 볼 때 시스템의 作用循環週期의 初期나 末期狀態만을 言及하는 경우 靜的狀態(static state)라 하며 出力이 다음 週期에 入力된다면 初期와 末期 사이를 動的狀態(dynamic state)라 한다. 動的狀態에서는 항상 屬性들의 값이 修正되게 되며 一般的으로 시스템이 不安定的이 되는 경우는 動的狀態에서의 修正値가 過多하거나 修正 자체가 많이 지연되는 때이다. 예를 들면 〔圖 2〕에서 보는 바와 같이 修正信號 發生時點에서 下向調整을 지시하지만 時間이 지연되어 이미 상당히 下向되었을 때 修正하게 되어 下限線을 지나게 된다.

〔圖 2〕 지연된 시스템修正



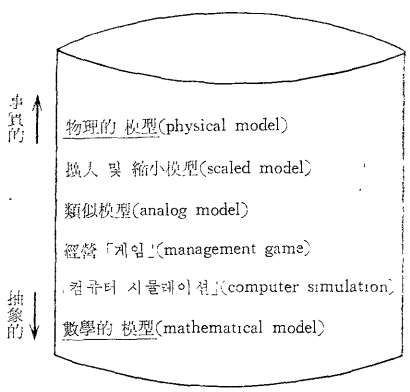
2. 模　　型

시스템을 硏究하기 위해서는 가능하면 實際 시스템으로 實驗을 하는 것이 가장 바람직하다. 그러나 이 방법은 불가능한 경우가 많고 또한 費用과 時間의 限界에 구애를 받는 경우가 대부분이다. 예를 들어 經濟시스템을 硏究하기 위하여 實際시스템에서 財貨의 供給과 需要를 임의로 바꾸면서 實驗을 할 수는 없는 것이다. 따라서 시스템의 硏究는 一般的으로 시스템의 模型(model)을 통하여 행하여지며 여기서 模型이란 시스템의 모든 상세한 부분을 고려하기에 앞서 그 시스템을 單純化 혹은 縮小化한 것을 말한다. 硏究의 目的에 따라 蒐集된 情報는 模型의 本體를 이루게 되므로 한 시스템에 대응하는 模型은 唯一하지 않고 따라서 模型設定者가 다르거나 혹은 同一人이라 하더라도 그의 見解에 따라 模型이 달라질 수 있다.

模型들이 實際시스템을 代置할 수는 없지만 硏究하는 사람이 적어도 關心 對象이 되는 特性들을 다룰 수 있게 模型을 設計하였다면 매우 有用한 것으로 보아야 한다. 또 간혹 未備한 점이 있다 하더라도 새로운 模型을 開發하는 데 필요한 情報를 提供하는 정도로도 그 價値를 찾을 수 있다. 따라서 模型들이 사실과 「같다」 혹은 「틀리다」라고 말할 수도 없고 말할 필요도 없다. 模型들의 價値는 이로써 나타내지는 시스템의 이해에 얼마나 도움을 주는가에 의해서 決定되는 것이다.

實際시스템의 現象을 觀察하고 假定들을 公式化하여 模型을 設定한 후에는 때때로 그 模

〔圖 3〕 模型의 區分



型을 評價 혹은 檢定하는 것이 필요하다. 이를 評價하는 첫째 條件은 우리가 알고 있는 모든 사실을 감안하여야 하고, 둘째로는 偏見이 없는 사람에 의해서 評價될 수 있는 豫測(prediction)能力이 있어야 한다. 실제의 評價作業은 새로운 觀測値가 얻어졌을 때 實際시스템과 模型에서 나온 결과의 一致程度를 알아봄으로써 遂行되며 현격한 차이가 나타날 때에는 模型이 修正되어야 한다.

이와 같은 점들을 감안해 볼 때 模型은 대체로 思考能力에 도움을 주기 위하여, 他人과의 意思疏通에 도움을 주기 위하여, 訓練 및 指針의 目的으로, 豫測手段으로, 實驗遂行에 도움을 주기 위하여 사용되는 등의 機能을 가졌다고 말할 수 있다.

模型들은 대응되는 본래의 시스템에 따라 구분되지만 〔圖 3〕과 같이 分類될 수도 있다.

物理的 模型이란 실제의 시스템元素들을 사용하는 模型으로서 예를 들어 自動車시스템의 硏究를 위해 같은 크기의 模型을 만드는 경우나 航空機나 船舶의 建造를 위한 縮小模型, 分子나 原子의 硏究를 위한 擴大模型 등이 이에 속한다. 類似模型은 實際시스템의 元素들의 性質과 비슷한 性質을 갖는 代替物質을 사용한 模型으로서, 「아나로그 컴퓨터」(analog computer)의 電壓이 財貨의 供給量을 表示한다든가 計算尺(slide rule) 등이 좋은 예이다. 또 圖表나 「그래프」 등도 時間, 數量, 나이 등을 나타내는 類似模型의 一種이라 할 수 있다. 〔圖 3〕의 위에서 아래로 내려올수록 人間과 「컴퓨터」가 서로 밀접한 연관을 갖게 되는데 그 중간적인 位置에 있는 것으로 經營「게임」이 있다. 軍隊의 指揮者나 經營者의 意思決定過程은 定型的인 형태로 模型化하기는 힘드나 흔히 訓練 등의 目的을 위해 「게임」이 사용된다. 이것의 특징은 人間이 「컴퓨터」와 相互作用하여 「컴퓨터」의 出力에 대하여 意思決定을 하고 이를 다시 「컴퓨터」에 入力하는 일련의 순환과정에 있다. 이것을 더욱 擴張시키면 전적으로 「컴퓨터」만을 사용하여 模型을 「시뮬레이션」할 수 있게 된다. 「컴퓨터 시뮬레이션」이란 「컴퓨터」를 이용하여 計算할 수 있는 「컴퓨터」模型을 設定하고 이를 통하여 시스템을 硏究하는 것을 意味한다. 이를 위해서는 一般的인 「컴퓨터」言語를 사용하여 「프로그램」을 作成하든가 아니면 特殊模型에 맞는 「시뮬레이션」言語를 사용하기도 한다. 끝으로 數學的 模型이란 物理的 元素를 사용하지 않고 數學的인 記號로서 시스템元素 및 屬性들을 表示하여 이를 變數 혹은 「파라미터」(parameter)로 사용하고 이들의 關係를 函數로 表示하는 模型을 말한다. 오늘날의 「컴퓨터」는 거의 모든 硏究에 必須不可缺한 要素이므로 앞으로 本稿에서는 數學的 模型이 「컴퓨터 시뮬레이션」을 포함하는 意味로 보기로 한다.

이상의 模型分類 이외에 靜的模型(static or cross-section model)과 動的模型(dynamic

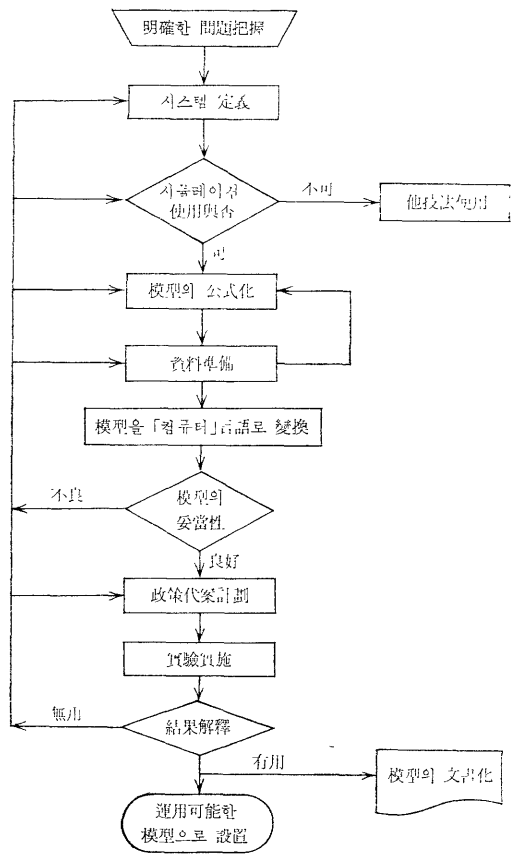
model or time-series model), 決定的模型(deterministic model)과 確率的模型(stochastic model), 離散模型(discrete model)과 連續模型(continuous model) 등으로 區分하기도 한다.

3. 시뮬레이션

複雜한 시스템의 設計나 運營을 맡은 사람들에게 가장 좋은 分析道具 중의 하나가 시뮬레이션(simulation)[1]이다. 이를 槪略的으로 定義하면 다음과 같다. 시뮬레이션이란 實際시스템을 模型化하고 이 模型을 통하여 시스템의 行動狀態를 이해함은 물론 시스템運營上의 여러 戰略的인 代案을 評價할 목적으로 실험을 행하는 것을 말한다. 위의 定義에 의하면 시뮬레이션을 廣義로 解釋하여 模型設定까지도 이에 포함시키고 있으나 이보다 더 좁은 意味로 시뮬레이션을 定義하는 경우도 많다. 또한 종이와 연필만 가지고 혹은 탁상計算器만으로도 시뮬레이션이 遂行될 수 있으므로 「컴퓨터 시뮬레이션」만이 시뮬레이션이라고 制限할 수도 없다.

시뮬레이션이 問題를 解決하는 데 좋은 方法이긴 하나 效果的이라고 말하기는 힘들다. 예를 들어 模型이 단순한 形態를 이루고 있고

〔圖 4〕 시뮬레이션의 段階



解析的인 方法으로 그 解答을 구할 수 있다면 시뮬레이션을 할 필요가 없다. 따라서 硏究하는 사람은 여러가지 가능한 分析道具를 사전에 檢討함은 물론 費用과 時間 등도 考慮對象에 포함시켜야 한다. 또한 어떤 問題에 대한 理解의 視角이 달라지고 資料가 追加되는 등의 變化가 생기면 시뮬레이션 사용의 妥當性을 再檢討해야만 한다. 한편 「컴퓨터 시뮬레이션」은 「컴퓨터」使用時間이 길고 標本數가 큰 경우가 많기 때문에 그 費用이 解析的으로 解答을 구하는 方法(가능할 때) 보다 비싼 경우가 있으나 應用性, 正確性 등의 이유로 해서 널리 사용되고 있다.

1) 確率的模型을 시뮬레이션할 때 Monte-Carlo 方法을 사용하는 경우가 있어 이를 시뮬레이션과 혼동하기 쉬우나, Monte-Carlo 方法은 1940年 後半에 von Neumann과 Ulan이 처음으로 사용하여 名稱한 것으로 어떠한 確率分布를 갖는 亂數를 生成하여 資料를 만드는 標本抽出方法에 사용하면가 다음과 같이 積分値를 推定하는 데 쓰이는 技法이다. 函數 $f(x)$에 대하여 解析的으로 積分 $\theta=\int_0^1 f(x)dx$가 存在하지 않을 때 (0,1)區間의 一樣分布로부터 $X_1, X_2, \cdots, X_n$의 標本을 만들어 $\theta=\frac{1}{n}\sum_{i=1}^{n} f(X_i)$를 計算하면 이것이 θ의 不偏推定量이 된다.

시뮬레이션技法의 段階는 〔圖 4〕에서도 볼 수 있는 바와 같이 여러 段階로 구분할 수 있다. 이 각각의 段階는 模型設定의 過程으로 생각할 수 있으며 이들 段階에 대한 說明은 다음과 같다.

어떠한 問題의 解答을 얻기 위해서는 명확한 問題把握이 必須的임은 다 아는 사실이다. 그러나 때때로 經營者層에서는 問題가 있다는 사실은 알면서도 問題가 무엇인지 정확히 把握하지 못하는 경우가 있다. 이러한 경우 시스템分析은 經營者의 指揮 아래 시스템의 槪略的인 것부터 研究를 시작하게 되며 追加情報나 代案이 생길 때마다 把握된 問題의 妥當性을 檢討한 후 필요에 따라 修正을 가함으로써 이루어진다. 이 過程 중에서 가장 중요한 것은 研究對象이 되는 시스템에 대한 定義를 내리는 일이며 모든 시스템은 더 큰 시스템의 部分시스템이기 때문에 制限條件이 또한 明示되어야 한다. 問題는 만족되지 않은 필요의 상태로 나타나고 이 상태는 시스템의 영향이 원하는 결과를 낳지 않을 때 發生한다. 따라서 問題는 다음과 같이 數學的으로 定義된다.

$P_t = |D_t - A_t|$

P_t : t時點의 問題

D_t : t時點의 원하는 狀態

A_t : t時點의 實際狀態

시스템을 定義하는 첫단계는 시스템과 시스템環境을 구분하는 일이며 이들 이외의 것들과도 구분해 놓아야 한다. 그렇지 않으면 무한히 많은 關係가 가능하게 되어 실제의 시스템을 簡略化한 模型을 構想할 수 없기 때문이다. 또 시스템元素, 關係 등이 너무 簡略化되어 있으면 시스템은 自明하게 되고 이와 반대로 시스템元素, 關係들이 너무 細細한 부분까지 포함하고 있으면 다루기 힘들 뿐만 아니라 費用도 많이 들게 된다. 모든 사람들은 거의 실제 시스템과 같을 정도로 자세히 시스템을 시뮬레이션하기를 원하나 그것은 그리 바람직하지 못하다. 왜냐하면 「컴퓨터 프로그램」의 어려움, 비싼 費用 등의 問題뿐만 아니라 실제로 너무 자세하고 複雜한 시스템을 사용할 경우 본래의 중요한 性質과 關係 등이 상실되어 정확한 模型의 認識이 어렵기 때문이다. 그러므로 問題에 대답을 줄 수 있을 정도로 혹은 研究의 目的에 적절한 정도로 시스템이 定義되어야 한다. 많은 경우에 제대로 問題把握이 되면 시뮬레이션을 할 필요도 없이 問題點이 解決되어 이 段階에서 끝날 수도 있다.

거의 모든 研究에는 資料準備가 필요하다. 그러므로 어떠한 資料가 필요한가, 資料가 研究目的에 妥當한가, 現存하는 資料가 있는가, 없는 資料는 어떻게 만드는가 등에 대한 判斷이 優先되어야 한다. 더우기 經驗的資料(empirical data)를 사용하느냐 혹은 亂數(random number)를 만들어 사용하느냐의 選擇은 중요하다. 經驗的資料를 사용할 경우의 分析은 過去의 行態가 未來에도 계속된다는 假定下에 이루어지게 되므로 將來의 行態에 대해 새로운 것을 말해 주지 못하며 단지 過去를 시뮬레이션하는 것에 불과하다. 그러나 만일 亂數를 사용할 수 있다면 「컴퓨터」를 이용할 수 있다는 점 이외에 必須的은 아니지만 入力資料의 결과에 대한 感度分析(sensitivity analysis)이 용이하다는 利點이 있다. 이러한 資料의 選擇과 아울러 資料의 妥當性, 그 形態, 理論的 分布 혹은 過去 行態에 대한 適合性 등은 理論的 見解를 떠나서 시뮬레이션의 成功與否에 매우 중요한 要素이다.

이와 같이 하여 資料의 準備가 完了되고 模型이 公式化되면 이를 「컴퓨터」 言語로 變換하여야 하는데 이 경우 어느 特殊한 시뮬레이션에 맞는 「소프트 웨어」를 사용할 수도 있고[2] FORTRAN, ALGOL, BASIC 등 일반적인 「컴퓨터」言語를 사용하여 直接 「프로그래밍」할 수도 있다.

模型이 실제 시스템에 정확한 혹은 맞는 模型이다 라고 端的으로 말한다는 것은 불가능한 일이다. 그러나 실제의 使用者 側面에서 보면 시뮬레이션으로부터 얻은 시스템의 推論方法에 대해 어느 정도의 信賴가 필요하며 이를 위한 方法이 模型에 대한 妥當性檢討이다. 가끔 시뮬레이션을 행하는 分析者가 그의 假定을 감춤으로써 상당히 위험한 誤謬를 내포하고 있는 결과가 그대로 默認되어 通過되는 경우도 있을 수 있으므로 模型의 妥當性檢討는 매우 중요한 過程이다. 妥當性檢討의 方法에는 經驗이 많은 사람 혹은 專門家들로 하여금 여러 가지로 수행된 시뮬레이션 結果를 判斷하도록 하는 方法과, 統計的 檢定(statistical test)이 많이 사용되는 시스템 假定에 대한 檢定 또는 시스템 入出力變換에 대한 檢定 등이 있다. 槪念的으로 볼 때 模型의 妥當性檢討란 模型이 실제 시스템과 동일한 形態로 行動하는가의 問題와 模型의 實驗을 통해서 얻어진 推論結果가 실제와 一致하는가를 檢討하는 것이라 할 수 있다. 그러나 經驗的 硏究에서는 標本數가 너무 적은 경우, 資料가 너무 統合(aggregation)된 경우, 資料가 부정확한 경우 등의 理由로 檢定 自體가 곤란한 때가 있다.

政策代案計劃이란 우리가 원하는 情報를 얻기 위하여 模型의 實驗을 어떻게 할 것인가를 計劃하는 것을 말하며 그 目的은 필요한 實驗試行回數의 最小化 및 硏究者의 知識習得過程을 提供하는 데 있다. 이는 廣義의 實驗計劃法(experimental design)에 속하며 均衡狀態에 도달하는 데 영향을 주는 初期條件 決定과 最小의 標本數로 分散이 적은 結果의 算出이 매우 중요한 要件이 된다. 初期條件이 適切하게 주어지면 均衡狀態에 도달하는 컴퓨터 時間을 減少시킬 수 있고, 많은 標本數를 사용하면 시뮬레이션에서 일어날 수 있는 여러 問題를 극복할 수 있으나 역시 費用이 많이 들 뿐 아니라 앞에서 言及한 바와 같이 標本數 자체에 制約이 있으므로 分散을 적게 하는 技法이 필요하게 되며, 이러한 計劃은 模型이 複雜할수록 더욱 重要性을 띠게 된다.

다음의 過程은 원하는 情報를 얻기 위하여 실제의 實驗 혹은 「컴퓨터」 作業을 遂行하는 節次이다. 이 作業에서 나온 結果로부터 模型의 弱點을 發見할 수 있으므로 우리의 硏究目的이 達成될 때까지 그 이전의 단계를 反復的으로 再點檢할 수 있다. 또한 시뮬레이션에서 중요한 槪念 중의 하나가 感度分析인바 이는 어느 限定된 범위 내에서 「파라미터」 값을 體系的으로 變動시켜 가면서 그 效果를 觀測하는 것이다. 微細한 「파라미터」 값의 變動에도 變數의 結果値가 크게 變動하는 경우는 이 「파라미터」를 더욱 정확히 推定하려는 노력이 필요하다는 것을 意味하게 된다.

마지막으로 모든 과정이 〔圖 4〕에서와 같이 反復的으로 施行되어 통과되면 模型의 細分化, 使用者의 訓鍊, 條件變化에 따른 調整, 結果의 信賴性 提高 등을 위하여 運用可能한 模型으

2) 一般的인 시뮬레이션 「소프트웨어」로는 GPSS, SIM SCRIPT, SIMULA, DINAMO 등이 있고 計量模型의 시뮬레이션에는 여러가지 統計 「패키지」가 있다 (第Ⅳ章 參照).

로 設置하여야 한다. 이 過程에 投入되는 時間은 보통 소홀히 되기 쉽지만 놀랍게도 시뮬레이션을 위해 投入되는 全體時間의 10~25%에 이르러야 한다고 報告되고 있다. 模型設置와 아울러 중요한 것은 模型의 文書化(documentation) 作業인데 詳細한 記錄은 模型의 運用 및 設置에 상당한 도움을 주게 된다. 더욱 記錄이 잘 되어 있으면 模型을 開發한 사람이 不在時에도 그 模型을 계속 사용할 수 있으며 또한 「컴퓨터 프로그램」 중의 많은 「서브프로그램」(subprogram)들은 장래의 다른 業務에도 사용이 가능하다.

이상의 시뮬레이션 過程은 하나의 技法이며 理論的 學問은 아니다. 그러므로 模型設定에 있어서 무엇이 포함되어야만 하고 또 무엇이 遂行되어야 한다는 뚜렷한 法則이 없이 많은 부분이 經驗에 依存하며 試行錯誤도 새로운 模型의 開發에 많은 情報를 주게 된다.

Ⅲ. 計量模型

1. 計量經濟模型

計量經濟模型[3](econometric model)은 一般的으로 數學的 模型이며 確率變數를 포함하고 있는 確率的 模型이다. 確率變數는 전형적인 형태가 加法的攪亂項(additive stochastic disturbance term)으로 나타나는데 그 理由들은 다음과 같이 說明될 수 있다.

實質的으로 모든 變數는 모든 變數의 函數로 看做되며 이를 變數의 關係는 函數 $Y=f(X_1, X_2, X_3, \cdots)$의 形態로 表示된다. 그러나 모든 變數들 전부를 模型說明變數로 택하는 것은 아니며 시스템의 定義에서 言及한 바와 같이 많은 變數들을 제외하거나 혹은 Y에 가장 영향을 주는 X들만을 고려한 函數, 즉 $Y=g(X_1)+h(X_2, X_3, \cdots)$로 보아 $h(X_2, X_3, \cdots)=u$로 간주하게 된다. 여기서 u는 제외된 變數들의 영향을 나타내며 平均이 0이고 有限한 分散을 갖는 確率變數로 假定한다. 이와 같은 方法으로 變數들의 關係를 函數로 나타낼 때 정확한 函數의 形態를 모르므로 假定된 函數形態와 실제의 關係 사이에 存在하는 誤差가 攪亂項으로서 나타나기 때문에 確率變數의 形態가 加法的攪亂項의 모습을 띠게 된다. 실제의 計量分析에서도 資料를 사용하여 分析한 후에 函數의 形態가 修正되는 경우가 자주 있다.

또 다른 理由로 變數의 測定誤差(errors in measurement)를 들 수 있는데 예를 들어 $Z=\alpha+\beta X$의 정확한 관계가 成立하고, Z 變數의 값을 測定할 때 誤差가 있어서 Y로 測定되거나 혹은 Z를 測定하기 불가능하여 대신 Y를 測定할 경우 $Y=Z+u$(u는 測定誤差)가 되어 $Y=\alpha+\beta X+u$의 式을 얻게 된다. 그러나 이러한 要因들은 複合的으로 發生할 수 있고 또한 加法的이기 때문에 攪亂項 하나로 表示될 수도 있다[4].

加法的攪亂項이 있는 線型計量經濟模型[5]은

3) 計量經濟模型은 社會科學 全般에 걸쳐 사용되기 때문에 計量模型이라 할 수 있으나 聯立方程式 體系로 說明하므로 구분하였음.

4) X의 測定誤差가 있을 때 한 攪亂項으로 나타나는 것은 非現實的이므로 이 경우는 「파라미터」 推定方法으로 IV(instrumental variable) 方法이나 ML(maximum likelihood) 方法 등을 사용함.

5) 計量經濟模型에서 線型이라 함은 方程式이 「파라미터」에 대하여 線型임을 意味하며 線型性을 假定하는 主要理由는 數學的, 統計學的 理論展開에 많은 利點이 있기 때문이다. 오늘날 非線型模型도 많이 쓰이고 있는데 이는 과거의 계산상 난점이 컴퓨터 「소프트웨어」

다음과 같은 m個의 式으로 表示되는 構造型(structural form)으로 構成된다.

$$\beta_{11}Y_1+\beta_{12}Y_2+\cdots\cdots+\beta_{1m}Y_m+\gamma_{11}X_1+\cdots\cdots+\gamma_{1k}X_k=u_1$$
$$\beta_{21}Y_1+\beta_{22}Y_2+\cdots\cdots+\beta_{2m}Y_m+\gamma_{21}X_1+\cdots\cdots+\gamma_{2k}X_k=u_2$$
$$\cdots\cdots\cdots\cdots$$
$$\beta_{m1}Y_1+\beta_{m2}Y_2+\cdots\cdots+\beta_{mm}Y_m+\gamma_{m1}X_1+\cdots+\gamma_{mk}X_k=u_m$$

$Y_1, \cdots\cdots, Y_m$: m個의 內生變數

$X_1, \cdots\cdots, X_k$: k個의 外生變數 혹은 lag付內生變數

$u_1, \cdots\cdots, u_m$: m個의 攪亂項

行列型으로는 $B\underset{\sim}{y}+\Gamma\underset{\sim}{x}=\underset{\sim}{u}$로 表示되며 $m\times m$ 行列 B는 正則行列(nonsingular matrix)로 假定한다. 이 行列式의 양변에 B^{-1}을 곱하면

$$\underset{\sim}{y}=-B^{-1}\Gamma\underset{\sim}{x}+B^{-1}\underset{\sim}{u}$$

혹은

$$\underset{\sim}{y}=\Pi\underset{\sim}{x}+\underset{\sim}{v}\quad(\Pi=-B^{-1}\Gamma,\ \underset{\sim}{v}=B^{-1}\underset{\sim}{u})$$

을 얻는데 이를 誘導型(reduced form)이라 한다.

攪亂項 確率벡타 $\underset{\sim}{u}$에 대한 統計的 假定은 「파라미터」推定에서 중요한 役割을 하며 이러한 假定은 실제로 資料를 사용할 때 반드시 再檢定되어야 한다. 왜냐하면 모순된 假定으로부터 얻은 推定値는 각 變數들간의 關係 즉, 方程式을 틀리게 決定짓기 때문이다.

構造型이 n個의 觀測値를 가진 行列式

$$B\underset{\sim}{y}_i+\Gamma\underset{\sim}{x}_i=\underset{\sim}{u}_i,\quad i=1,2,\cdots\cdots,n$$

와 같이 表示될 때 다음과 같이 전형적인 $\underset{\sim}{u}$에 대한 假定을 한다.

$$E(\underset{\sim}{u}_i)=\underset{\sim}{0},\quad i=1,2,\cdots\cdots,n$$
$$Cov(\underset{\sim}{u}_i,\underset{\sim}{u}_j)=E(\underset{\sim}{u}_i\underset{\sim}{u}_j')=\begin{cases}\Sigma & i=j\text{일 때},\\ 0 & i\neq j\text{일 때}\end{cases}\quad i=1,2,\cdots,n$$

더욱 $\underset{\sim}{u}$가 多變量正規分布(multivariate normal distribution)를 하는 것으로 假定하는 것이 一般的인데 이는 中心極限定理(central limit theorem)를 應用한 것으로 보아 무난한 것이다. 또 Σ가 항상 對角行列이 되지 않는 것은 聯立方程式模型이 單一方程式模型과 다르다는 점을 意味하며 각 式에 個別的으로 單一方程式模型의 推定方式을 適用하지 않는 理由가 되는 것이다[6].

實際로 構造型의 「파라미터」推定은 誘導型의 推定 즉, Π의 推定과

$$\Omega=Cov(\underset{\sim}{v})=E(\underset{\sim}{v}\underset{\sim}{v}')=B^{-1}E(\underset{\sim}{u}\underset{\sim}{u}')(B^{-1})'$$
$$=B^{-1}\Sigma(B^{-1})'$$

의 推定으로 하게 되는데 이때 이들 結果로부터 본래 構造式에서의 B, Γ, Σ의 推定値를 얻을 수 있는가의 識別問題(identification problem)가 대두된다.

그러나 $\Pi=-B^{-1}\Gamma$와 $\Omega=B^{-1}\Sigma(B^{-1})'$의 關係만을 가지고는 B, Γ, Σ가 決定될 수 없기 때문에 構造型에서의 制限條件이 誘導型의 推定에 事前的으로 감안되어야 한다. 대개 構造型의 制限條件은 「파라미터」가 0이라든가 線型結合이 일정한 값을 갖는다든가 하는 條件 등으로 주어지며, 마찬가지로 Σ의 어떤 元素가 0이라는 條件을 들 수 있다.

지금까지 聯立方程式(simultaneous equat-

와 數値解析의 발달로 극복되어졌고 그 推定理論이 발달되었기 때문이다.

6) 線型聯立方程式에서 Σ가 對角行列이거나, Σ가 對角行列이 아니더라도 誘導型의 각 式의 說明變數가 同一할 때는 각 式에 個別的으로 單一方程式 最小自乘法을 사용할 수 있다.

ions)으로 構成된 模型을 說明하였으나 많은 경우에 있어서 單一方程式(single equation)으로 模型이 設定된다. 다음 節에서는 이에 대하여 여러 가지 혼돈하기 쉬운 槪念과 模型區分에 관하여 說明하고자 한다.

2. 一般線型模型과 線型回歸模型

單一方程式 線型模型은 定量模型(quantitative model)과 定性模型(qualitative model)으로 나눌 수 있으며 定量模型에는 一般線型模型(general linear model: GLM)과 線型回歸模型(linear regression model: LRM)이 있고 定性模型에는 計劃模型(design model: DM)과 分散成分模型(components-of-variance model: CVM) 등이 있다. 이러한 模型들은 서로 밀접하게 연관되어 있고 推論하는 過程 또한 매우 유사하다.

GLM은 線型模型의 代表的인 模型이며 다른 세 가지 模型도 GLM의 變化된 形態로 볼 수 있으므로 우선 GLM부터 定義하기로 한다.

(GLM)

$Y=\mu(x)+\varepsilon$

Y : 從屬變數(觀測可能한 確率變數)

x : 獨立變數의 벡타(確率變數가 아님)

ε : 觀測할 수 없는 確率變數

$\mu(x)$: 未知의 「파라미터」의 線型函數로서 函數形態는 알고 있음

위의 式에 더하여 ε의 平均, 分散, 分布函數 등이 포함된 確率的 假定이 함께 模型의 일부를 이룬다.

또한 $\mu(x)$의 函數形態는 다음과 같이 k個의 未知의 「파라미터」 $\beta_1, \beta_2, \cdots\cdots, \beta_k$의 線型函數로서 表示된다.

$$\mu(x)=\beta_1 q_1(x)+\beta_2 q_2(x)+\cdots\cdots+\beta_k q_k(x)$$

$q_i(x)$: 未知의 「파라미터」가 없는 x의 函數(函數形態는 알고 있음)

이와 같은 GLM은 母集團 變數들의 관계를 說明한 것으로서 β_i들과 ε의 分散 등의 「파라미터」들을 推定키 위해서 母集團으로부터 n個의 標本을 抽出하여 다음과 같이 表示된다.

$$Y_i=u(x_i)+\varepsilon_i,\quad i=1, 2, \cdots\cdots, n$$

이와 같이 n個의 標本을 抽出한 式으로 表示된 模型을 標本模型(sample model)이라고 하고 앞의 模型 $Y=\mu(x)+\varepsilon$을 母集團模型(population model)이라 한다.

$\mu(x)$의 가장 전형적인 形態는 $\mu(x)=\sum_{j=1}^{k}\beta_j x_j$로서 標本模型으로 表示하면

$$Y_i=\sum_{j=1}^{k}x_{ij}\beta_j+\varepsilon_i$$

가 되고 行列表示法을 사용하면 다음의 形態가 된다.

$$\underset{\sim}{Y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$$

$\underset{\sim}{Y}$: $n\times 1$ 確率벡타(觀測可能함)

X : $n\times k$ 行列(X의 元素는 確率變數가 아님)

$\underset{\sim}{\beta}$: $k\times 1$ 「파라미터」벡타 ($\underset{\sim}{\beta}\varepsilon\Omega_\beta$: 「파라미터」 空間)

$\underset{\sim}{\varepsilon}$: $n\times 1$ 確率벡타(觀測할 수 없음)

$E(\underset{\sim}{\varepsilon})=\underset{\sim}{0},\ Cov(\underset{\sim}{\varepsilon})=\Sigma$

여기서 $x_{i1}=1,\ i=1, 2, \cdots, n$이면 $Y_i=\beta_1+\sum_{j=2}^{k}x_{ij}\beta_j+\varepsilon_i$가 되어 切片(intercept)이 있는 模型이 되며 더욱 $k=2$일 때 즉, $Y_i=\beta_1+\beta_2 X_i+\varepsilon_i$인 경우 이 模型을 單純線型(標本)模型(simple linear model)이라 부른다. 또 이 模型의

$\underset{\sim}{\varepsilon}$에 대하여 正規分布를 한다는 假定을 더 주기도 하고 Σ의 形態에 새로운 假定을 追加하기도 한다[7]. 한편 「파라미터」空間 Ω_β는 보통 k次元 Euclidean空間이나 이 空間이 制限되면 즉, $\underset{\sim}{\beta}$에 制限條件이 들어가면 制限된 模型(restricted model)이라고 한다.

LRM(linear regression model: 線型回歸模型)[8]은 GLM과 많이 혼돈하게 되는 模型으로 GLM과 매우 유사한 形態를 띠고 있다. 두 模型間의 가장 중요한 差異點은 GLM에서의 獨立變數는 確率變數가 아니지만 LRM에서의 獨立變數는 確率變數라는 것이다. 그러므로 GLM에서는 「파라미터」 자체를 推定하지만 LRM에서는 $(Z|\underset{\sim}{X}=\underset{\sim}{x})$의 條件附分布에 나타나는 「파라미터」를 推定하게 된다. 이를 자세히 說明하기 위해 두 變數 Z와 X가 二變量正規分布 $N(\underset{\sim}{\mu}, \Sigma)$를 한다고 假定하면

$$\underset{\sim}{\mu}=\begin{pmatrix}\mu_Z\\ \mu_X\end{pmatrix},\quad \Sigma=\begin{pmatrix}\sigma_Z^2 & \sigma_{ZX}\\ \sigma_{ZX} & \sigma_X^2\end{pmatrix},\quad \rho=\frac{\sigma_{ZX}}{\sqrt{\sigma_Z^2\sigma_X^2}}$$

이고 二變量正規分布의 性質에 의해서

$$E(Z|X=x)=\mu_Z+\left(\frac{\sigma_{ZX}}{\sigma_X^2}\right)(x-\mu_X)=\beta_1+\beta_2 x$$
$$Var(Z|X=x)=\sigma_Z^2(1-\rho^2)=\sigma^2$$

이 된다. 또 確率變數 $Y=(Z|X=x)$의 경우

$$Y=(Z|X=x)=\mu_Y(x)+\varepsilon=\beta_1+\beta_2 x+\varepsilon$$
$$\varepsilon\sim N(0,\sigma^2)$$

이 되어 GLM과 유사한 形態을 갖는다[9].

예를 들어 서울에서 30살 때의 所得이 x_0인 男子의 40살 때 所得이 얼마인가를 알아 보는 問題를 생각해 보자. 40살 때의 所得을 Z, 30살 때의 所得을 X라 하고 두 變數가 二變量正規分布를 한다고 假定하면 어떤 사람의 40살 때 所得의 豫測値는 μ_Z로 놓을 수 있다. 왜냐하면 서울에 40살 男子의 所得은 $N(\mu_Z, \sigma_Z^2)$分布를 하기 때문이다. 그러나 이 사람의 30살 때 所得 x_0를 안다면 μ_Z 보다 좋은 豫測値를 얻을 수 있다. 그 理由는 이 사람의 40세 때 所得이 $Y=(Z|X=x_0)$의 分布를 따르므로

$$\sigma_Y^2=\sigma^2=(1-\rho^2)\sigma_Z^2\leqslant\sigma_Z^2$$

이 成立되기 때문이다. 따라서 豫測値로 μ_Z보다는 $\mu_Y(x_0)=\beta_1+\beta_2 x_0$를 사용하게 된다. 다시 말해서 30세 때 x_0의 所得인 男子의 경우 μ_Z보다는 $\mu_Y(x_0)$를 40세 때의 所得으로 갖을 確率이 크다는 것이다. 물론 여기서 μ_Z도 모르고 $\mu_Y(x_0)$도 未知의 「파라미터」를 갖기 때문에 우리는 n個의 標本(Z_1, X_1), (Z_2, X_2), ……, (Z_n, X_n)을 抽出한 후에 μ_Z, μ_X, σ_Z^2, σ_X^2, σ_{ZX}를 推定하여 β_1, β_2, σ^2의 推定値를 얻게 된다.

이제 LRM을 定義하면 다음과 같다.

(LRM)

$k+1$個의 確率變數$(Z, X_1, X_2, \cdots\cdots, X_k)$가 平均 μ와 共分散行列 Σ를 갖는 分布를 하고

$$E(Z|X_1=x_1,\ X_2=x_2, \cdots\cdots,\ X_k=x_k)$$
$$=\mu_Y(x_1, x_2, \cdots\cdots, x_k)$$
$$=\sum_{i=0}^{k}\beta_i q_i(x_1, x_2, \cdots\cdots, x_k),$$
$$Y=(Z|X_1=x_1,\ X_2=x_2,\ \cdots, X_k=x_k)$$
$$=(Z|X=\underset{\sim}{x})$$

---

7) $\underset{\sim}{\varepsilon}$에 대하여 確率的 假定이 전혀 없을 때 즉, $\underset{\sim}{Y}=X\underset{\sim}{\beta}+\underset{\sim}{d}$, $\underset{\sim}{d}$는 非確率的 誤差일 때 最小自乘模型(least square model)이 되며, $\underset{\sim}{\varepsilon}$이 確率벡타이고 $\Sigma=\sigma^2 I$이면 Gauss-Markov模型, $\Sigma=\sigma^2 H$(H는 正値對稱行列)이면 Aitken 模型이라 부른다. 물론 $E(\varepsilon_i)=0$, $\mathrm{Var}(\varepsilon_i)=0$이면 決定的模型임을 알 수 있다.

8) 回歸(regression)라는 말은 英國의 優生學者 F.Galton (1822-1911)이 처음 使用한 것으로 아버지의 身長 x와 아들의 身長 y의 關係式 $y_i=\alpha+\beta x_i$에서 $\beta<1$임을 간파하고 이 直線은 點$(\bar{x}, \bar{y})$을 지나고 $x<\bar{x}$이면 $y>x$이고 $x>\bar{x}$이면 $y<x$이 되어 아들들의 身長은 人間 全體의 平均身長에 되돌아 가려는 傾向이 있다는 사실로부터 由來되었다.

9) $\mu_Y(x)$가 항상 「파라미터」에 대하여 線型일 수는 없다. 그때는 非線型回歸模型(nonlinear regression model)이라 부른다.

$q_i(x)$: 未知의 「파라미터」가 없는 x의 函數(函數形態는 알고 있음)

$$Var(Z|X_1=x_1, X_2=x_2, \cdots\cdots, X_k=x_k) = Var(Y)=\sigma^2(\text{常數})$$

일 때 LRM은 다음의 式으로 表示된다.

$$Y=\mu_Y(x_1, x_2, \cdots\cdots, x_k)+\varepsilon, \quad E(\varepsilon)=0, \quad Var(Y)=\sigma^2$$

위의 定義에서 $q_0(x_1, \cdots\cdots, x_k)=1$ 이고 $q_i(x_1, x_2, \cdots, x_k)=x_i$, $i=1, 2, \cdots, k$일 때 이를 多重(線型) 回歸模型(multiple linear regression model)[10]이라 하며 실제로 가장 많이 쓰이는 模型이다. 또한 GLM에서와 마찬가지로 ε에 대한 分布의 假定이 함께 模型의 일부분을 이루게 된다. 보통 ε의 分布가 正規分布임을 假定하게 되는데 이는 더욱 세분되어 $(Z, X_1, X_2, \cdots, X_k)$가 多變量正規分布를 할 경우와 $(Z|X_1=x_1, \cdots, X_k=x_k)$만 正規分布를 하고 $(X_1, \cdots, X_k)$의 分布는 모르는 경우의 두 가지로 나누어진다[11].

다음으로 定性模型의 計劃模型(design model: DM)을 說明하기로 한다. 定性模型이나 定量模型이란 말은 獨立變數의 性格에 의해 구분되며 만약 獨立變數가 金額, 時間, 重量, 길이 등 計測할 수 있는 것이면 定量變數(quantitative variable)라 하고 種類, 色, 性, 生産方法 등 어떤 特性을 나타내는 것이면 定性變數(qualitative variable 혹은 categorical variable)라 한다. 一般的으로 變數를 區分하는 方法은 〈表 1〉과 같다.

특히 GLM에서 獨立變數들이 定性變數이고 從屬變數가 連續變數일 때 計劃模型이 되며 이를 ANOVA型 模型(ANOVA type model)이라고도 부른다. 이 模型을 예를 들어 說明하면, 어떤 工場에서 A와 B의 두 가지 방법에 의해 製品을 生産한다고 했을 때 각 方法에 따라 두 母集團($i=1, 2$)이 있고 각 母集團에서 세개의 標本을 抽出한 후 强度를 實驗하여 얻은 값을 Y_{i1}, Y_{i2}, Y_{i3}, $i=1$(A方法), $i=2$(B方法)이라고 하면 模型은 다음과 같이 表示된다.

$$Y_{11}=\mu_1+\varepsilon_{11}$$
$$Y_{12}=\mu_1+\varepsilon_{12}$$

〈表 1〉 變數의 一般的인 區分

二元變數(dichotomous variable) 예) Yes · No, 男 · 女 無順多元變數(nonordered polytomous variable) 예) 다섯개의 다른 機械들 順序多元變數(ordered polytomous variable) 예) 高所得層, 中産層, 低所得層	定性變數 (qualitative or categorical variable)	離散變數 (discrete variable)
整數値變數(integer-valued variable) 예) 나이 連續變數(continuous variable)		

10) 函數 $E(Z|X=x)=\mu_Y(x)$가 線型函數이라 함은 $\mu_Y(x)$가 x에 대하여 線型임을 말한다. 즉, 註 5)와 註 9)에서의 線型, 非線型의 구분은 「파라미터」의 函數로 본 것이나 線型回歸函數에서의 "線型"이란 말은 x에 대한 線型을 意味한다.

11) 두 경우의 「파라미터」 推定에서 ML(maximum likelihood) 推定値는 同一함.

$$Y_{13}=\mu_1+\varepsilon_{13}$$
$$Y_{21}=\mu_2+\varepsilon_{21}$$
$$Y_{22}=\mu_2+\varepsilon_{22}$$
$$Y_{23}=\mu_2+\varepsilon_{23}$$

(μ_1, σ_1^2) : A 母集團分布의 平均과 分散

(μ_2, σ_2^2) : B 母集團分布의 平均과 分散

혹은

$$\underset{\sim}{Y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$$

$$\underset{\sim}{Y}=\begin{bmatrix}Y_{11}\\Y_{12}\\Y_{13}\\Y_{21}\\Y_{22}\\Y_{23}\end{bmatrix},\quad X=\begin{bmatrix}1&0\\1&0\\1&0\\0&1\\0&1\\0&1\end{bmatrix},$$

$$\underset{\sim}{\beta}=\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\quad \underset{\sim}{\varepsilon}=\begin{bmatrix}\varepsilon_{11}\\\varepsilon_{12}\\\varepsilon_{13}\\\varepsilon_{21}\\\varepsilon_{22}\\\varepsilon_{23}\end{bmatrix}$$

여기서 X의 元素는 0 혹은 1 만이 나타나며 이는 각 式에서 μ_i의 有無를 나타내 준다. 한편 이 形態는 GLM과 같기 때문에 GLM에 適用되는 推定方法이 그대로 適用될 수 있다.

위의 예에서 $\mu_i=\mu+\alpha_i$로 놓는다면 μ는 전체 標本의 平均强度를 나타내게 되고 α_i는 生産方法에 따라 감안되어야 할 强度를 表示하게 된다(α_i는 −부호를 가질 수 있다). 이때의 模型은

$$Y_{ij}=\mu+\alpha_i+\varepsilon_{ij},\quad i=1,2,\quad j=1,2,3$$

혹은

$$\underset{\sim}{Y}=X^*\underset{\sim}{\beta}^*+\underset{\sim}{\varepsilon}$$

$$\underset{\sim}{Y}=\begin{bmatrix}Y_{11}\\Y_{12}\\Y_{13}\\Y_{21}\\Y_{22}\\Y_{23}\end{bmatrix},\quad X^*=\begin{bmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&0&1\end{bmatrix},$$

$$\underset{\sim}{\beta}^*=\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\end{bmatrix},\quad \underset{\sim}{\varepsilon}=\begin{bmatrix}\varepsilon_{11}\\\varepsilon_{12}\\\varepsilon_{13}\\\varepsilon_{21}\\\varepsilon_{22}\\\varepsilon_{23}\end{bmatrix}$$

이 되어 실질적으로 같은 現象을 模型化한 것이 된다. 여기서 앞의 X의 階數는 列數와 같으므로 full rank 模型이라 하나 이와는 달리 X^*의 階數는 列數와 같지 않다[12].

(DM)

GLM $\underset{\sim}{Y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$에서 $n\times k$ 行列 X의 階數(rank) p가 $n>k\geqslant p$이고 어느 特別한 計劃에 의하여 X의 元素가 0 혹은 1 만을 갖는다면 이를 計劃模型(DM)이라 한다.

一般的으로 GLM에서는 X의 階數가 k이나[13] DM에서는 k보다 적다. 그러므로 DM에서는 特別한 推定方法을 사용하여야 한다[14].

보통 $\underset{\sim}{\varepsilon}$에 대한 統計的 假定은 다음의 두 가지 경우이다.

$$\underset{\sim}{\varepsilon}\sim N(\underset{\sim}{0},\ \sigma^2 I)\ \text{혹은}\ \underset{\sim}{\varepsilon}\sim(\underset{\sim}{0},\ \sigma^2 I)$$

앞의 DM模型 $Y_{ij}=\mu+\alpha_i+\varepsilon_{ij}$에서 α_i가 分散 σ_α^2을 갖는 確率變數일 때 이 模型은 分散成分模型(components-of-variance model: CVM)이 된다. 分散成分模型(CVM)을 實例를 들어 說明하면 다음과 같다.

어떤 機械製造會社에서 機械性能을 알아 보기 위하여 임의로 I個의 機械를 選擇한 후 J일 동안 實驗하여 Y_{ij}, $i=1, 2, \cdots, I$, $j=1, 2, \cdots,$

12) 計量經濟學에서의 dummy variable이 이러한 定性變數로서 同一한 性質을 갖음.

13) X의 階數가 k일지라도 多重共線性이 存在時에는 다른 推定方法을 사용함.

14) 여기서 特殊한 推定方法이란 計算上의 問題點으로서 一般化逆行列(generalized inverse matrix)를 사용한 計算方法임.

J를 測定하였다. Y_{ij}의 觀測値에 차이가 생기는 理由로 I個의 機械가 비록 같은 製造會社의 同一種類의 機械라도 그 性能이 다르다는 점과 實驗日이 다르다는 점의 두 가지를 들 수 있다. 그러므로 模型은

$$Y_{ij}=\mu+T_i+\varepsilon_{ij},\ i=1,2,\cdots,I,\ j=1,2,\cdots,J$$

로 表示되고, 機械에 따라 效果가 달라짐을 나타내는 確率變數인 T_i는 $E(T_i)=0$, $Var(T_i)=\sigma_T^2$을 갖고 i機械의 j번째 날의 效果를 나타내는 確率變數 ε_{ij}는 $E(\varepsilon_{ij})=0$, $Var(\varepsilon_{ij})=\sigma_\varepsilon^2$을 가지므로 우리의 目的은 결국 σ_T^2과 σ_ε^2을 推定하는 것이 된다. 즉,

$$\sigma_Y^2=\sigma_T^2+\sigma_\varepsilon^2$$

이 되어 確率變數 Y의 分散은 그 成分인 σ_T^2과 σ_ε^2으로 分解된다. 이때 T와 ε은 서로 相關되지 않는다고 假定하며 경우에 따라서는 그 分布函數가 追加되어 假定되기도 한다. 이를 行列로 表示하면 $I=2$, $J=3$일 때

$$\underset{\sim}{Y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$$

$$\underset{\sim}{Y}=\begin{pmatrix}Y_{11}\\Y_{12}\\Y_{13}\\Y_{21}\\Y_{22}\\Y_{23}\end{pmatrix},\quad X=\begin{pmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&0&1\end{pmatrix},$$

$$\underset{\sim}{\beta}=\begin{pmatrix}\mu\\\alpha\end{pmatrix}=\begin{pmatrix}\mu\\T_1\\T_2\end{pmatrix},\quad \underset{\sim}{\varepsilon}=\begin{pmatrix}\varepsilon_{11}\\\varepsilon_{12}\\\varepsilon_{13}\\\varepsilon_{21}\\\varepsilon_{22}\\\varepsilon_{23}\end{pmatrix}$$

와 같이 되며 이는 앞에서 言及한 DM과 같은 形態를 갖고 있음을 알 수 있다. 一般的으로 分散成分模型[15](CVM)은 다음과 같이 定義된다.

(CVM)

$\underset{\sim}{Y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$

$\underset{\sim}{Y}: n\times 1$ 確率벡타(觀測可能함)

$X=[\underset{\sim}{1}, X_1, X_2, \cdots, X_q]: n\times k$ 計劃行列

$X_i: n\times p_i$ 行列, $\sum_{i=1}^{q} p_i=k-1$

$\underset{\sim}{\beta}'=[\mu, \underset{\sim}{\beta}_1', \cdots, \underset{\sim}{\beta}_q']$, $\underset{\sim}{\beta}_i: p_i\times 1$ 確率벡타, μ와 $\underset{\sim}{\beta}_i$는 觀測不可能

$\underset{\sim}{\varepsilon}: n\times 1$ 確率벡타(觀測不可能)

$\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, \cdots, \underset{\sim}{\beta}_q$, $\underset{\sim}{\varepsilon}$은 서로 無相關

$E(\underset{\sim}{\beta}_i)=\underset{\sim}{0}$, $Cov(\underset{\sim}{\beta}_i)=\sigma_i^2 I$, $\sigma_i^2>0$, $i=1,2,\cdots,q$

$E(\underset{\sim}{\varepsilon})=\underset{\sim}{0}$, $Cov(\underset{\sim}{\varepsilon})=\sigma_\varepsilon^2 I$, $\sigma_\varepsilon^2>0$

이제까지 說明한 네 가지 전형적인 模型 이외에도 DM과 CVM이 혼합된 混合模型(mixed model)이 있고 그밖에 GLM과 DM, LRM과 DM의 混合模型 등이 있다.

Ⅳ. 統計「소프트웨어」

模型을 設定하고 시뮬레이션하기 위하여 특수한 시뮬레이션 言語를 사용하거나 혹은 직접 「프로그래밍」할 수 있음을 第Ⅱ章 第3節에서 이미 言及하였다. 그러나 實際의 많은 硏究分野에서는 確率的 模型을 사용하기 때문에 「파라미터」 推定 등을 위하여 統計學的 技法이 요구되며 따라서 이를 위한 컴퓨터 「소프트 웨어」가 다양하게 開發되어 왔다.

「프로그램 팩키지」(program package)란 使用者가 쓰기 편리하게 提供되는 컴퓨터 「소프

15) 統計學에서 DM을 Model I 혹은 fixed effect 模型이라 하며 CVM을 Model Ⅱ 혹은 random effect 模型이라 부른다.

트 웨어」로서 過去에는 특별한 分析에만 사용할 수 있는 「서브프로그램」(subprogram)들을 모아 놓은 「팩키지」가 많이 사용되었으나 최근에는 한번의 資料入力으로 여러 種類의 分析을 할 수 있는 「팩키지」가 널리 이용되고 있다. 前者는 대개의 경우 使用者의 「프로그램」에서 불러서 사용할 수 있으며 「프로그램」의 管理가 比較的 용이하다. 典型的이고 많이 쓰이는 「팩키지」로는 IMSL(International Mathematical and Statistical Libraries)과 SSP (Scientific Subroutine Package)를 들 수 있다. 한편 後者의 경우는 컴퓨터에 관한 知識이 없는 사람이라도 사용하기가 쉬우므로 使用者의 층이 넓고 資料加工能力 및 融通性이 높은 편이다. SAS(Statistical Analysis System), SPSS(Statistical Package for Social Science), BMDP(Biomedical Package) 등이 이에 속한다. 이 밖에도 많은 「팩키지」들이 여러 分野에서 널리 사용되고 있다.

이러한 각 「팩키지」를 評價하는 主眼點은 使用者나 使用目的에 따라 다르지만 대체로 사용의 便利性, 正確性, 擴張性, 文書化 程度, 資料構造, 出力形態, 圖表作成, 多樣性, 컴퓨터 機種에 대한 獨立性, 設置 및 管理의 容易性 등을 들 수 있다.

다음은 여러 「팩키지」들을 主 使用目的에 따라 구별하여 놓은 것으로 세계적으로 널리 쓰이는 「팩키지」들이며 比較的 優秀한 것들에 대해서는 간단한 說明을 追加하였다.

1. Data Management

가. **SIR** : SPSS와 BMDP를 직접 連結시켜 使用할 수 있는 「데이터 베이스」(Data Base) 管理시스템으로서 여러가지 優秀한 機能을 가지고 있음.

나. **MARK IV** : 報告書 및 圖表를 만들 뿐만 아니라 DB「파일」(file)을 만들거나 維持하는 機能을 갖고 있음.

다. **RAPID** : 「센서스」나 調査資料들로부터 DB를 만들고 管理하는 機能을 갖고 있으며 SPSS「파일」과 연결시켜 쓸 수 있음.

라. 기타 : UPDATE, ADABAS, RIQS, DATA3, FILEBOL, KPSIM/KPVER 등

2. Editing

가. **GES** : 原「프로그램」(source program)을 修正하지 않고도 設問資料 등을 쉽게 編輯, 校正할 수 있음.

나. **CONCOR** : 面談이나 設問調査의 構造를 把握하기 위해서 만든 「프로그램」임.

다. 기타 : CHARO, EDITCK, VCP-LCP 등

3. Tabulation

가. **PERSEE**: 4次元까지의 cross tabulation과 breakdown 등 各種 統計値의 table을 無制限으로 만들 수 있음.

나. **COCENTS** : 간단한 「파일」 뿐만 아니라 複雜하고 多段階的인 「파일」을 操作할 수 있으며 임의로 加重値를 줄 수도 있음. 또한 grouping이나 re-ordering에 의해서 새로운 變數를 만들 수 있고 간단한 統計値를 포함시킬 수 있음.

다. 기타 : SYNTAX II, LEDA, CYBER GENINT, ISIS, VDBS, GTS, CRESCAT,

CENT III, CENTS-AID, FILE 2.0, OSIRIS 2.4, GEN SUMMARY, TPL, RGSP, NCHS-XTAB, SURVEY, SAP 등

4. Survey Variance-Estimation

가. **HES VAR X-TAB** : Balanced half-sample replication 技法으로 分散을 計算하며 여러가지 多樣한 結果値를 볼 수 있음.

나. **SUPER CARP** : 共分散行列 및 回歸係數, 部分集團의 平均 및 合計 등과 回歸方程式에서의 各種 統計値와 誤差를 計算할 수 있는 등 여러가지 機能이 있음.

다. 기타 : SPLITHALF, MULTI-FRAME, MWD VARIANCE, GSS EST, CLFS VAR-COVAR, CESVP, KEYFITZ, CLUSTERS, GENVAR 등.

5. Survey Analysis

가. **EASYTRIEVE** : file maintenance, information retrieval, report writing을 위해서 만들어졌으며 複雜한 DB管理도 可能함.

나. **SURVEYOR/SURVENT** : 電算에 대한 經驗이 전혀 없는 사람도 使用할 수 있도록 만들어졌으며 batch로 혹은 interactive하게 使用할 수 있음.

다. 기타 : BTFSS, EXPRESS, FOCUS, DATAPLOT, PACKAGE X, SCSS, SPSS, SOUPAC, DATATEXT, P-STAT 78, DAS, SPMS, OSIRIS IV 등

6. General Statistical Programs

가. **SAS** : information storage and retrieval, data modification, programming, report writing, statistical analysis, file handling 등 各種 機能을 골고루 갖춘 매우 優秀한 「팩키지」임

나. **BMDP** : 各種 統計分析機能을 갖추고 있어 널리 使用되고 있으며 大型이나 小型 컴퓨터에서 使用이 可能함.

다. **NISAN** : 最近에 日本에서 새로이 開發된 方法에 의해 만들어진 것으로 「모델」을 만들 때 統計學을 잘 모르는 初步者들도 쉽게 使用할 수 있는 interactive 統計「팩키지」임.

라. **SPEAK EASY III** : 一般的인 目的이나 높은 水準의 統計分析을 위하여 interactive하게 만들어졌으며, 使用이 간편함.

마. **TROLL** : 統計分析을 할 수 있는 거의 모든 機能을 갖추고 있으며 특히 시뮬레이션을 수행하는 데 優秀한 機能을 갖추고 있음.

바. 기타 : CS, OMNITAB 80, HP STAT PACKS, MINITAB II, GENSTAT, IDA 등

7. Specific Statistical Analysis-Interactive

가. **GLIM** : 一般線型模型(GLM)을 위하여 만들어진 interactive 「팩키지」임.

나. **RUMMAGE** : 變數가 連續變數이거나

離散變數이든, 또 模型이 均衡計劃模型이거나 不均衡計劃模型이든가에 關係없이 모든 線型模型을 풀 수 있는 「팩키지」임.

다. 기타 : ISA, ISP, CMU-DAP, CADA, SURVO, AUTOGRP+, FORALL, AQD, STP, STATPAK, STATUTIL, MICROSTAT, A-STAT 등

8. Specific Statistical Analysis-Batch

가. **LSML 76** : 線型模型과 混合模型을 分析하기 위해서 만들어짐.

나. **LINWOOD/NONLINWOOD** : 線型 및 非線型模型을 處理할 수 있음.

다. 기타 : AMANCE, MAC/STAT, REG, TPD, ACPBCTET, MULPRES, STATSPLINE, ALLOC, POPAN, CAPTURE 등

9. Multi-Way Contingency Table Analysis

가. **GUHA** : 多次元 定性資料(multidimensional categorical data)의 統計處理를 위한 「팩키지」임.

나. **ECTA** : 段階的인 log-linear模型을 處理하기 위한 多樣한 機能을 갖고 있는 「팩키지」임.

다. 기타 : CATFIT, TAB-APL, MULTIQUAL, C-TAB, MLLSA 등.

10. Econometrics and Time Series

가. **TSP/DATATRAN** : 時系列分析을 위한 거의 모든 機能을 갖추고 있는 優秀한 「팩키지」임.

나. **B34S** : 여러가지 時系列分析用 機能을 갖고 있어(특히 Markov, probability model을 포함한 非線型推定이 가능)널리 使用되는 「프로그램」임.

다. 기타 : PACK, SHAZAM, TSP,QUAIL, KEIS/ORACLE 등

11. Mathematical Subroutine Libraries

가. **DATAPAC** : I/O를 free-format으로 處理할 수 있고, graphical analysis 및 data editing, sorting, ranking, deleting, subset extraction에 「서브루틴」(subroutine)을 提供하여 주는 「프로그램」들임.

나. **IMSL** : 統計와 數學分野를 망라한 500개 以上의 計算用 「서브루틴」(subroutine)으로 構成된 優秀한 「프로그램」이다.

다. 기타 : REPOMAT, NMGS2, NAG, LIBRARY, EISPACK 등

V. 結　　言

最近 社會科學의 모든 分野에서는 計量模型이 普遍化되고 이를 통한 硏究結果가 자주 發表되고 있으며 실제로 많이 사용되고 있다. 그러나 模型設定時 硏究對象이 되는 시스템에 대한 정확한 理解, 硏究目的이 되는 問題의 분명한 把握, 信憑性있는 資料의 뒷받침, 올바

른 分析方法의 選擇 등이 先行되지 않고는 그 結果에 대한 信賴度 提高를 期待하기는 어렵다. 특히 一般線型模型과 線型回歸模型 등의 分析時에 模型의 假定들에 대한 정확한 理解없이 단지 分析方法에만 의존한다든가 혹은 國內의 特殊한 制約與件을 감안하지 않은 채 外國의 模型을 그대로 模倣하려는 分析態度 등은 많은 危險性을 내포하게 된다. 따라서 위와 같은 여러 先決要件들을 충분히 고려한 후에 計量模型을 設定하여야만 그 有用性이 높아지며 많은 硏究者의 經驗 또한 蓄積될 수 있는 것이다.

한편 統計學을 이용한 資料의 分析이나 模型의 시뮬레이션은 정확한 理論이 存在하지 않을 때 어떤 推論을 誘導하기 위하여 사용되는 方法이기는 하나 항상 最先의 方法은 아니며 때로는 經驗論的 知識이 많은 寄與를 할 때도 있다. 未來의 美國人口를 豫測하는 수단으로 어떤 模型이 設定되었을 때 이 模型이 韓國의 人口豫測에 適用되어 一般化될 수 없는 경우, 標本數가 制限된 資料를 토대로 만들어진 模型이 그 範圍가 넓고 資料가 없는 부분의 性質을 效果的으로 說明하지 못하는 경우, 因果關係를 확실히 把握하지 않고 模型 自體만으로 모든 現象을 說明하려는 경우 등은 模型을 濫用하고 있는 좋은 예라고 볼 수 있다.

또한 近來에 많은 大學校와 硏究所 등에서 컴퓨터를 이용한 模型分析이 急增하고 있으나 이에 필요한 「소프트웨어」의 確保 및 컴퓨터 사용에 많은 苦痛이 뒤따르고 있다. 그러므로 國內에서도 이러한 未備點들을 補完하려는 노력이 시급하며 이 같은 노력의 결과로 넓은 층의 經驗者들을 배출함은 물론 社會科學分野의 發展에도 寄與하는 바가 매우 크리라 생각된다. 本稿에서 다룬 몇가지 중요한 槪念들과 模型에 대한 구분 및 「소프트웨어」에 대한 紹介는 이러한 發展에 미약하나마 도움이 되리라 期待된다.

▷ 參 考 文 獻 ◁

高麗大學校 貿易硏究所, 「經濟 및 經營政策樹立을 위한 管理技法과 컴퓨터 應用」, 西獨 Mannheim大學校 產業經營硏究所 System Research Group 아시아地域巡廻세미나, 1974.

郭相瓊, 『計量經濟學』, 茶山出版社, 1982.

朴聖炫, 『回歸分析』, 大英社, 1983.

呂運邦, 『計量分析의 電算處理』, 韓國開發硏究院, 1977.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W., *Discrete Multivariate Analysis*, The MIT Press, 1975.

Dhrymes, P.J., *Econometrics*, Harper & Row, Publishers, 1970.

Francis, I., *Statistical Software*, North Holland,1981.

Goldberger, A.S., *Econometric Theory*, John Wiley & Sons, Inc., 1964.

Gordon, G., *System Simulation*, Prentice-Hall, Inc., 1969.

Graybill, F.A., *Theory and Application of the Linear Model*, Duxbury Press, 1976.

Green, P.E., *Analyzing Multivariate Data*, The Dryden Press, 1978.

Griliches, Z. and Intriligator, M.D., *Handbook of Econometrics*, North Holland, 1983.

Gunst, R.F. and Mason, R.L., *Regression Analysis and Its Application*, Marcel Dekker, Inc., 1980.

Hopeman, R.J., *Systems Analysis and Operations Management*, Charles E. Merrill Publishing Co., 1969.

Johnston, J., *Econometric Methods*, McGraw-Hill, Ltd., 1972.

Kennedy, W.J. and Gentle, J.E., *Statistical Computing*, Marcel Dekker, Inc., 1980.

Kmenta, J., *Elements of Econometrics*, Macmillan Publishing Co., Inc., 1971.

McMillan, C. and Gonzalez, R.F., *Systems Analysis*, Richard D. Irwin, Inc., 1968.

Opter, S.L., *Systems Analysis for Business and Industrial Problem Solving*, Prentice-Hall, Inc., 1965.

Searle, S.R., *Linear Models*, John Wiley & Sons, Inc., 1971.

Shannon, R.E., *Systems Simulation*, Prentice-Hall, Inc., 1975.

Raj, B. and Ullah, A., *Econometrics*, Croom Helm Ltd., 1981.

Rivett, P., *Principles of Model Building*, John Wiley & Sons, Inc., 1972.

Yeo, W.B., *"Fitting Seemingly Unrelated Nonlinear Regression"*, Thesis for M.S., Iowa State University, 1980.