aw-Hill Inc., 1969.

Robert W. Miller, Schedule, Cost, and Profit Control with PERT, McGraw-Hill Inc., 1963.

論 文

〈國 內〉

孔大植 : Network Analysis를 適用한 工程管理, 經營論集, 第Ⅱ卷 第1號, 1968년 3月, 韓國經營研究所(서울大學校 商科大學)

蔣孝健 : PERT·CPM手法에 의한 設計工程管理研究, 高麗大學校 大學院, 1970년.

金用憲 : Critical Path Method의 基礎理論과 그 導入可能性에 대한 一考察, 經營論集, 第Ⅱ卷 第2號, 1966년 7월, 韓國經營研究所(서울大校 商科大學)

鄭福圭 : PERT技法의 基本原理와 導入適用에 관한 考察, 產業經濟 第4輯, 1970년, 產業經濟研究所(嶺南大學校 商經大學)

鄭福圭 : PERT·CPM技法에 의한 Slack Time處理와 Line Balancing, 經營論叢 第9輯, 1973년, 經營研究所(嶺南大學校 商經大學).

鄭福圭 : CPM技法에 의한 日程短縮과 費用節減方法의 兩立可能性에 관한 實際的 考察, 嶺南大學校 論文集, 1972년.

鄭福圭 : PERT/man Power Control, 商經學報 第三輯, 嶺南大學校, 1972년.

〈國 外〉

Warren Dusenbury, CPM for New Product Introductions, Harvard Business Review, July-August, 1964.

W. Miller, How to Plan and Control with PERT, Harvard Business Review, March-April 1962(Vol. 40, No. 2)

E. B. Berman, Resource Allocation in a PERT Network under Continuous Activity Time-Cost Functions, Management Science, July 1964 (Vol. 10, No. 4)

James E. Kelly. Jr., "Critical Path Planning and Scheduling: Mathematical Basis", Operations Research, May-June 1961.

F. Klevy, G. L. Thompson & J. D. Wiest, The ABCs of The Critical Path Method, Harvard Business Review, Sep-Oct., 1963.

J. J. Moder, "How Do CPM Scheduling without a Computer", Engineering Newsrecord, March, 1964.

W. R. Ross, PERT/cost Resource Allocation Procedure, The Accounting Review, American Accounting Association, July, 1966.

其他 參考資料

PERT·CPM制度 Manual, 韓國產業開發研究所, 1970년 7월.

PERT·CPM制度化 方案 調査研究, 附屬資料, 韓國產業開發研究所, 1970년 7월.

PERT·CPM制度導入 및 施行을 위한 調査研究報告書 및 附屬回表, 1969년 11월, 韓國產業開發研究所.

工程狀況報告 第65號, 1972년 11월 20일 P製鐵株式會社.

設備別 基本建設工程計劃, 1970년 8월 12일 P製鐵株式會社.

韓國OR學會誌
第5卷 第1號
1980年 4月

회귀분석을 이용한 Data Editing

허 문 열

원래 다량자료의 정리(Large Scale Data Screening)는 어떤 정립된 이론에 의해 수행되는 것보다는 그 자료자체가 가지는 성격과 자원의 Availability 등을 고려하여 수행되는 것이 상례이다. 여기서는 여러 방법 중 자료가 모두 수치로 나타나는 경우 자료정리의 한 유용한 방법으로 회귀분석을 사용하는 방법에 대해 설명코자 한다.

테이블식 자료

데이타의 양이 많아지는 경우 여러 이유에서 깨끗한 결과를 맞나기 힘들며 그 이유를 들면

1) 현장에서 자료 수집하는 사람들의 오류
2) 수집된 자료의 컴퓨터 입력에 있어서의 오류
3) 자료운반시 혹은 보관시 일부 망실로 사후 조정에 의한 오류
4) 고의적인 자료의 변환

등으로 볼 수가 있겠다.

보통의 경우 자료의 형태는 테이블 형태로 표시할 수 있으며 실제 모든 자료가 테이블 형태로 보관되는 것이 효과적이라는 점은 많은 학자와 경험자들에 의해 주장되어 온 바이다. 이렇게 테이블 형태로 보관된 자료는 다음과 같이 보여질 수 있다.

순서 \ 성질	X_1 X_2…… X_p	Y_1 Y_2…… Y_m
1		
2		
3		
⋮		
⋮		
n		

(테이블 1)

이상의 테이블은 n개의 자료수와 각자료에 대해 크게 두 가지의 특성이 주어졌다. 즉 주성분 특성(X_1, X_2, …… X_p)과 부성분 특성(Y_1, Y_2…… Y_m)으로써 전체는 $n\times(p+m)$개의 Item이 있다고 하자. 대개의 경우 각각 자료에서 처음 몇 개의 항목은 Key가 되는 성분으로 신중히 처리될 뿐 아니라 변화가 없으나 그 외의 부성분 특성에 해당되는 값들이 자주 변하게 되며 오류를 만들어 내는 경우가 많겠다.

따라서 부성분 특성에 대한 자료의 정리를 회귀분석을 통하여 정리하고자 한다.

회귀 분석

회귀분석의 모델은 다음과 같이 나타날 수 있다.

$$y=b_0+b_1X_1+b_2X_2+\cdots\cdots+b_pX_p+e$$

여기서, $X_1, X_2, \cdots\cdots, X_p$는 p개의 주어진 입력이며 y는 이러한 입력으로 발생되는 출력으로 나타낼 수 있다. 그리고 e는 이 모델로 설명을 할 수 없는 어떤 요소이며 보통 평균은 0이며 분산은 σ^2로써 나타난다. 물론 σ^2는 알 수 없는 값이며 회귀분석 결과 추정치가 얻어질 수 있다. 또 $b_0, b_1, \cdots\cdots, b_p$는 회귀분석후 얻어져야 하는 미지의 상수이다. 회귀분석을 한 결과 σ^2, $b_0, b_1, b_2, \cdots\cdots b_p$의 추정치를 $S^2, a_0, a_1, \cdots\cdots a_p$라고 하고 모델에 이 추정치를 대입하여 얻어낸 출력을 $\hat{y}$라고 하면

$$r=y-\hat{y}$$

에 의해 실제치와 추정치의 잔여치(Residual) r이 구해진다. 또한 이 잔여치는 평균이 0이고 자료의 수가 많을 경우 분산이 σ^2에 가까워진다. 정규분포를 가정하면 잔여치가 $\pm 2S$ (S^2는 σ^2의 추정치로써 얻어진다)를 초과하는 비율이 거의 0%에 가까웁게 된다. 또 하나 자료정리에서 많이 나타나는 회귀분석의 하나로는 Dummy Variable의 기법이다. 이는 예를 들어 남·여 성별인 경우(1, 0) 혹은 (1, 2) 혹은 (M,F)로 나타나며 이를 회귀분석에 적용키 위하여는 Dummy Variable X_1, X_2를 다음과 같이 만들어 낸다.

$X_1=0, X_2=1$, if Male

$X_1=1$, $X_2=0$, if Female

따라서 남자인 경우 X_2만 모델에 나타나며 여자인 경우 X_1만 모델에 나타나게 된다. Dummy Variable을 모델에 첨가하는 편이 모델을 따로 만들어 내는 것보다 효율적인 경우가 나타나는 점은 다음과 같은 예에서 분명하다. 전체의 표본이 500명이고 남자가 250,

* 韓國科學技術究研所 電算開發센터

여자가 250인 경우 Dummy를 쓰면 표본수가 500으로 모델을 만들어 낼 수 있으나 남여 구별하여 사용할 경우 표본수가 절반인 250 줄어드는 점이겠다. 그러나 주의를 해야 할 점은 Dummy를 사용할 경우 설명 변수들이 남여에 동일한 의미로써 설명력을 가정함으로 경우에 따라서 사용되어야 하겠다. 또한 이미 설치된 여러가지의 통계 Package를 사용할 경우 상수부분과의 연관에 신경을 써야 하겠다.

자료 정리

회귀분석을 이용하여 테이블 1에 있는 각각의 부성분 특성($Y_1, \cdots\cdots Y_m$)에 대해 잔여치($r_1, r_2, \cdots\cdots r_m$)을 구했다고 하자. 이 잔여치 중에서 절대치가 예를 들어 $|2S_i|, i=1, \cdots\cdots m$를 초과하는 항목만 뽑아서 Print 하여 다음과 같은 테이블을 만듬으로써 부성분 특성들의 정리를 조직적으로 할 수 있다.

순서 \ 성질	$X_1, \cdots\cdots X_p$	$r_1, r_2, \cdots\cdots r_m$
1		
2		
⋮		
n		

(테이블 2)

또한 경우에 따라 실측치에 잔여치를 조정하여 추정치를 원래 자료에 삽입함으로써 깨끗한 자료를 만들 수 있다. 만약 추정치를 실측치에 대체시키는 경우가 불가능한 경우라 할지라도 잔여치의 크기를 봄으로써 그 자료가 얼마나 깨끗한지 아닌지를 쉽게 발견해 낼 수 있겠다. 회귀분석을 하기 전에 손쉬운 자료정리를 위해서 또한 회귀분석이 올바로 수행되기 위해서 먼저 Descriptive Statistics(평균, 편차, 최빈치, 중간치, Percentile 등——SAS인 경우 Univariate로써 얻어진다)를 조사하는 것이 바람직하다. 마지막으로 재미있는 예제를 들고자 한다. 다음은 성인 여성 30명에 대한 신장, 몸무게, 가슴둘레가 mm, kg, mm로 테이블 3과 같이 나타나 있다.

이 자료는 모두 정리가 되있으며 주성분 특성을 신장, 몸무게라 하고, 부성분 특성을 가슴둘레라 하면

$$(\text{가슴둘레}) = b_0 + b_1 \cdot (\text{신장}) + b_2 \cdot (\text{몸무게}) + e$$

로써 모델을 만들 수가 있겠다. 이러한 경우 만약 30번째 사람의 자료가 다음과 같은 3가지 경우에 틀리게 정리되어 있다고 하자.

1) 신장이 15500으로 2) 몸무게가 5300으로

3) 가슴둘레가 8500으로

표본번호	키	가슴둘레	몸무게
1	1611	940	570
2	1650	850	520
3	1647	940	540
4	1632	810	520
5	1626	934	620
6	1617	851	570
7	1615	962	500
8	1609	875	500
9	1603	850	520
10	1602	891	560
11	1584	870	570
12	1534	780	450
13	1580	870	520
14	1595	852	500
15	1625	834	530
16	1670	950	660
17	1685	900	620
18	1630	807	460
19	1560	771	470
20	1536	805	480
21	1507	790	450
22	1571	815	500
23	1568	880	520
24	1565	852	520
25	1562	835	510
26	1556	809	520
27	1554	865	480
28	1550	832	500
29	1550	775	450
30	1550	858	530

(테이블 3)

이런 경우에 각각의 회귀분석의 R^2, 상수부분, 그리고 신장, 몸무게에 대한 계수, 잔여치의 분산(S_i^2)의 추정치는 각 경우별로 테이블 4와 같이 나타나며 이 결과는 독자들이 연구할 가치가 충분히 있으므로 음미하기를 바란다.

	$R^2(\%)$	상수부분	키	몸무게	S_i^2
오류가 없는경우	51.37%	112.01	0.29	0.51	1333
1)	47.97%	493.07	0.0001	0.68	1427
2)	39.23%	−326.45	0.73	0.008	1666
3)	7.5%	15859	−12.05	8.48	1979750

(테이블 4)