정보학연구 제3권 제1호 2000년 3월(pp.37-44)
The Studies of Information Technology, Vol.3, No.1

다이아몬드 구매가격 예측을 위한 통계적 단순 선형회기 최적화 모형에 관한 연구

이영욱*

요 약

다이아몬드 구매 예측 가격은 캐럿, 색깔, 투명도, 품질등급, 절단상태 및 캐럿 당 $ 가격의 6가지 요소에 의하여 영향을 받는다. 본 연구의 목적은 이러한 구매가격을 예측하기 위한 선형 회기모형을 구하고 이를 통계적 방법으로 검증하는데 있다. 최적화 모형은 부적격 검정결과 통계적으로 적합성을 갖는 $\hat{y} = 102 / (-1.5575 + 0.3099 \log x) + \varepsilon$ 의 단순 회기모형으로 정규 분포성, 등분산성 및 대칭성의 특성을 갖는다.

Ⅰ. 서론

1950년대부터 다이아몬드는 여인들의 마음을 사로잡는 가장 인기 있는 보석중의 하나로 여겨왔다. 대개 구매자인 여인들의 구매력과 기호품중의 하나로 각광을 받아온 것이 사실이며 이를 팔기 위한 다이아몬드 도매상들의 투자대상의 보석이기도 하였다. 그러나 보석은 중량뿐만 아니라 색깔, 투명도, 품질 등에 따라 가격이 천차만별이다.

본 연구의 목적은 99개의 다이아몬드 데이터를 기초로 다이아몬드의 특성에 따른 가격에 대한 예측 모형을 통계적으로 분석, 개발하고 가장 적절한 모형에 따른 가격을 예측하기 위한 것이다. 이를 위한 통계적 연구로 가장 많이 사용되고 있는 가격 대 캐럿에 관한 선형회귀 방정식에 의한 모형을 구한 후 모형의 적절성을 분석하였으며 모형의 정규 분포성, 분산성 및 오차 특성에 따른 선형 최적화 모형을 제시하였다.

이러한 연구결과를 바탕으로 앞으로 무한한 가능성을 지닌 전자상거래 시장에서 예측하기 어려운 다이아몬드의 구매예측 가격을 제시할 수 있는 연구방안을 마련할 수 있을 것이다.

각 다이아몬드는 6개의 특성 즉 캐럿(Carat), 색깔(Color), 투명도(Clarity), 품질 등급(Cert), 절단상태(Cut) 및 가격(Price)의 변수로 기술할 수 있다.[4] 캐럿(Carat)은 다이아몬드의 중량(또는 크기)을 나타낸다. 색깔(Color)은 D에서 J까지 코드 화하여 나타내는데 D는 완전히 백색인 것, E와 F는 무색인 것, G에서 J 까지는 거의 무색인 다이아몬드를 나타낸다. 그 중에서 I는 보석 품질 중 최상품에 해당한다. 품질은 상품(higher quality)에 속한 다이아몬드이어야 투자와 구매의 대상이 된다. 투명도(Clarity)는 보석의 내부 또는 표면상에 흠집의 정도를 나타낸다. 즉, 얼마나 맑고 투명한가를 나타내는 척도가 된다. 투명도 F는 내부 흠집을 가지고 있음을 나타내며 VVS1과 VVS2는 내부에 육안으로 식별할 수 있는 매우 약간의 흠집을 갖고 있음

* 세명대학교 컴퓨터과학과 부교수

을 나타낸다. SI1과 SI2는 육안으로는 식별할 수 없는 약간의 흠집을 갖고 있는 것을 말한다. 보통 SI1부터 보석으로 간주하며 투자 품질의 대상이 된다. 품질 등급(Cert)은 보석의 등급을 매기는 기구에 의한 인증 품질을 나타낸다. G는 미국 보석협회(Gemological Institute of America)에 의한 인증 품질을 나타내는 데 가장 유명한 등급 판정기구에 의한 품질을 나타낸다. O는 기타 협회를 나타내며 N은 중량, 색깔, 투명도, 절단상태 등의 자격등급을 갖고 있지 않은 품질의 다이아몬드를 나타낸다. 가격에 영향을 미치는 다섯 번째 변수는 절단 상태(Cut)이다. 절단 상태는 4개의 코드로 나타내는데 A는 절단 상태가 완전한 것을 나타내며 B와 C는 A보다는 못한 것, N은 전문적인 절단 등급으로 간주하지 않는 경우를 말한다. 가격(Price)은 $/캐럿의 단위로 다이아몬드의 구매가격을 나타낸다.

본 논문은 제1장의 서론과 제2장 가격 대 캐럿에 관한 단순 선형 회기분석 모형에서 중량 대비 가격에 관한 적절한 모형을 구하고 통계적 방법에 의한 검정절차를 거쳐 제시한 모형 방정식이 유효함을 검증하였다. 제3장에서는 변수의 변형에 의한 예측 모형의 적절성을 분석하였고 제4장에서는 최적화 모형을 구하고 이러한 최적화 모형을 통하여 캐럿-중량에 의한 적절한 구매예측 가격을 제시하였으며 제5장에서 결론 및 향후 연구방향을 언급하였다.

II. 가격 대 캐럿 단순 선형 회기분석 모형

가격 대 캐럿(중량)에 관한 단순 선형 회기분석 모형은 종속변수 또는 반응변수 y(=price)와 독립변수 x(=carat) 사이에 선형관계가 있다고 가정하는 경우

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.1)$$

로 표현할 수 있다.[1]

여기서 β_0와 β_1은 추정하여야 할 계수이며 ε은 기대값 0과 분산 σ^2을 갖는 오차항으로 나타낸다. (그림 1)은 데이터 파일의 일부를 나타내고 있으며 전체 데이터 파일은 99개의 다이아몬드 데이터를 포함하고 있다.

OBS	CARAT	COLOR	CLARITY	CERT	CUT	PRICE
1	0.36	E	VS1	G	C	2139
2	0.43	E	VVS2	G	C	2516
3	0.46	G	IF	G	A	2284
4	0.53	D	VVS1	O	B	4635
5	0.52	E	VVS1	O	A	4645
6	0.51	E	VS1	G	C	3148
7	0.58	F	VVS2	O	A	3245
8	0.50	F	VS2	O	A	2600
9	0.59	G	VS1	O	A	2485
10	0.55	G	VS2	O	C	2516
11	0.55	H	VVS1	G	B	2866
12	0.53	H	VS1	O	A	2344
13	0.59	I	VVS2	O	A	2185
14	0.60	E	VVS1	O	B	4172
15	0.68	F	VVS1	G	A	3914
16	0.64	G	VS1	G	C	2678
17	0.69	J	VS1	N	N	1672
18	0.73	E	VVS1	G	A	4862
19	0.71	F	IF	O	B	4429
20	0.70	G	VVS1	G	A	4252
21	0.78	G	VS1	O	A	2781
22	0.70	H	IF	G	B	3605
23	0.79	I	IF	G	B	3214
24	0.80	E	IF	G	A	5253
25	0.87	F	IF	G	A	4841
26	0.84	H	IF	G	B	4202
27	0.92	D	VVS1	G	C	5459
28	0.94	E	VVS1	G	A	5665
29	0.97	E	VVS2	G	C	4996
30	0.91	F	VVS2	G	C	3966

(그림 1) 다이아몬드 데이터 파일의 일부

위 데이터 파일로부터 (2.1)식에 관한 단순 선형 회귀분석 모형은 SAS(Statistical Analysis

System) 출력으로부터

$$\hat{y} = 3614.71 + 1044.85x + \varepsilon \qquad (2.2)$$

의 식을 얻는다.

위 단순선형 회기분석 방정식에 관한 검정절차는 다음과 같다.

검정절차:

(i) 귀무 가설(Alternatives); Ho : $\beta_1 = 0$, Ha : $\beta_1 \neq 0$

(ii) 판정 규칙(Decision Rule);

① 만일 $F^* \leq F(1-\alpha, 1, n-2)$ 이면, Ho로 결정

② 만일 $F^* > F(1-\alpha, 1, n-2)$ 이면, Ha로 결정

위 (2.2)식에 관한 컴퓨터 출력으로부터 $F^* = 26.299 > F(1-\alpha, 1, 97)$ 이므로 $\beta_1 \neq 0$인 Ha로 결정한다. 이것은 캐럿 x와 가격 y 사이에 선형 관계가 있음이 유효하다는 것을 나타내며 캐럿 x와 가격 y에 관한 단순선형 회기분석 방정식이 +의 기울기를 갖는 증가함수 관계에 있음을 나타낸다. 그러나 이에 관한 상관관계를 나타내는 R^2 값은 R^2=0.2133 (상관계수 r= +0.4618)으로 낮은 값을 나타내고 있으며 그래프 또한 선형모형이 아니므로 캐럿 x와 가격 y에 관한 관계를 예측하기가 곤란하다. 따라서 통계적인 방법을 사용하여 이를 적절한 선형모형으로 변형시켜 재분석하는 것이 필요하다.

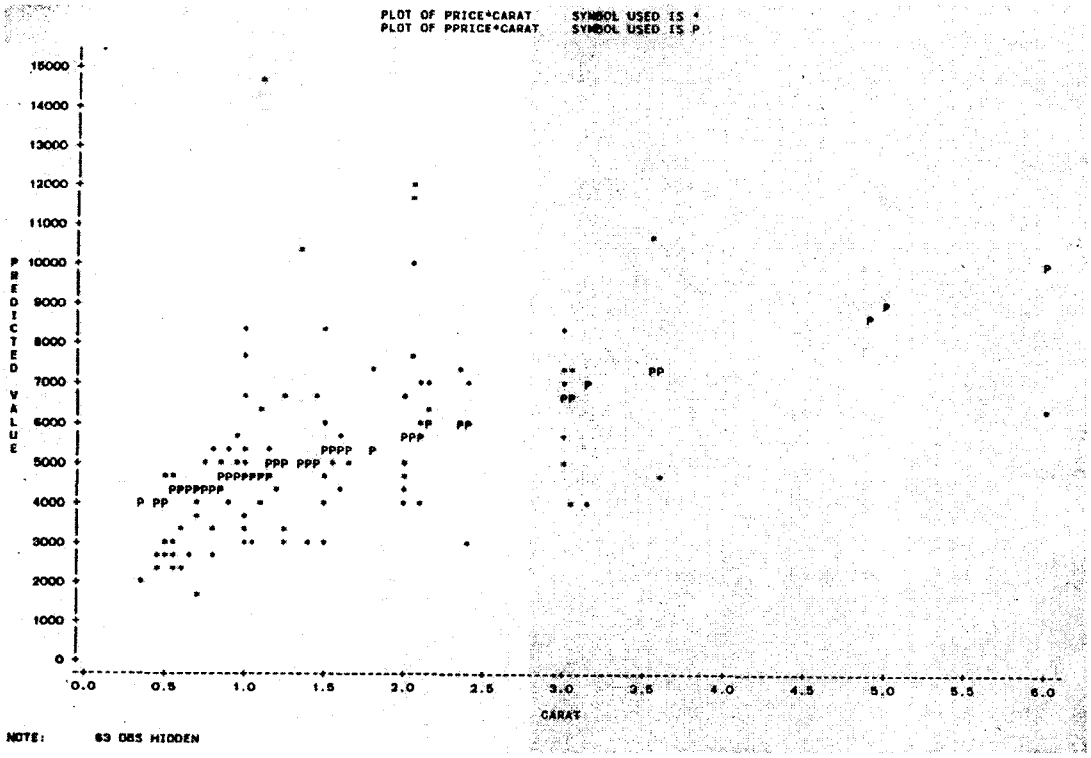
Ⅲ. 모형 적절성 (Model Adequacy) 분석

(그림 2)에서 알 수 있는 바와 같이 대부분의 데이터들의 분포특성은 증가하는 분포를 보여주고 있으며 비선형 분포임을 알 수 있다.

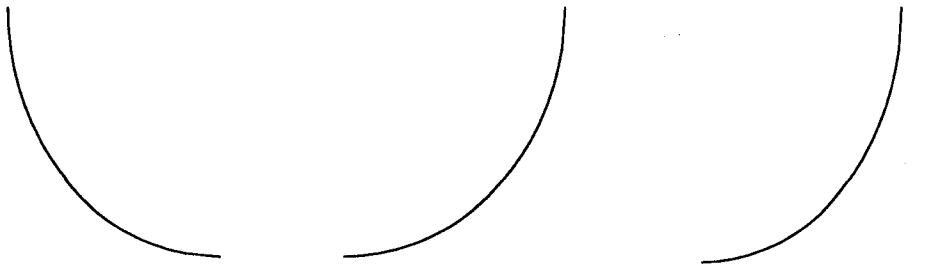
따라서 오차변수에 관한 특성도 비선형 분포로 증가하는 분포를 보여준다. 그러나 통계적 분석결과는 선형관계를 보여주고 있다. 따라서 통계적인 방법에 의하여 가격 예측을 가능케 하는 어떤 적절한 선형 분포를 찾아야 할 것이다. 이제 데이터의 선형화를 위한 모형의 변형을 고려해 보자.

3.1 비선형 분포에 관한 개선

다음 (그림 3)의 일반적 비선형 분포 중에서 선형분포에 접근하는 두 개의 분포 logx와 $\sqrt{x}$를 고려하여 두 분포결과를 비교해 보면, logx가 보다 더 선형에 가깝다고 볼 수 있다. (그림 4)는 두 분포도에 관하여 오차를 나타내는 잔차(Residual)에 대한 예측치(Prohibits)와의 비교 분포도이다.



(그림 2) 다이아몬드 데이터의 증가하는 비선형 분포

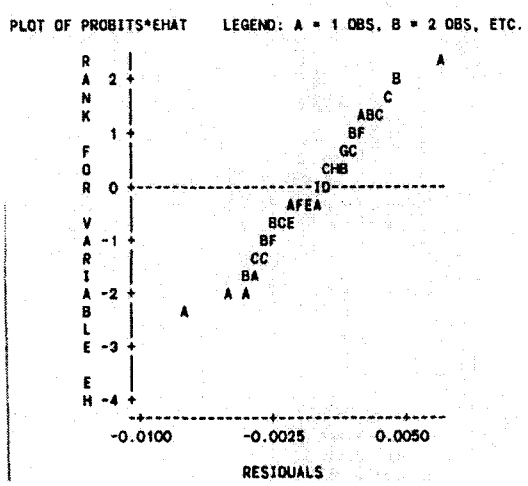


(a) $x' = \log x$ 또는 $x' = \sqrt{x}$

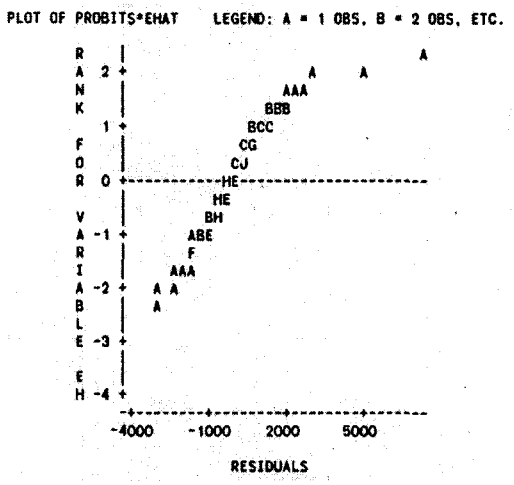
(b) $x' = x^2$ 또는 $x' = \exp(x)$

(c) $x' = 1/x$ 또는 $x' = \exp(-x)$

(그림 3) 전형적인 비선형 분포



(a) logx의 예



(2) √x의 예

(그림 4) logx와 √x의 잔차에 대한 예측치 비교 분포도

3.2 부등 분산(Unequal Variance) 분포에 관한 개선

이미 독립변수 x(=log carat)에 관한 변형모형을 구했으므로 종속변수 y(=price)에 관하여 적절한 선형분포 모형을 찾아보자. 이제 부등 분산(Unequal Variance) 분포에 관한 개선된

모형을 구하기 위하여는 선형성(Linearity), 대칭성(Symmetry) 및 등분산성(Constant Variance) 등을 고려한 최적 데이터를 찾아야 한다. 일반적 부등 분산 분포 중에서 선형분포에 접근하는 분포 1/y = logx의 변형 모형이 식(2.2)의 원래의 데이터 분포보다 더 선형에 가깝다고 볼 수 있다. <표 1>은 두 분포에 관한 통계적 분석결과를 나타낸다.

<표 1>에서 알 수 있는 바와 같이 유효검정을 위한 F와 상관관계를 나타내는 R^2의 값이 변형모형에서 많이 향상되었고 오차항에 관한 SSE/SSTO 값은 감소되었다. 따라서 원하는 분포특성이 향상되었음을 알 수 있다. 위 변형모형의 분석결과는 아직 불만족스런 상태이므로 또 하나의 변형모형 -1/ √ ^y = logx를 시도해 보자.

컴퓨터 출력으로부터 -1/ √ ^y = logx의 변형모형(^y은 y의 기대치) 은 선형특성뿐만 아니라 등분산성 및 대칭성을 갖고 있으나 여전히 표준화 잔차 분포로부터 알 수 있는 바와 같이 이상치(Outlier)를 갖고 있음을 알 수 있다. 따라서 다음 <표 2>는 이러한 이상치를 제거한 경우의 분석 결과를 나타내고 있다. <표 2>로부터 이상치가 제거된 경우에는 분석하기에 보다 좋은 데이터를 제공하게 된다. 이러한 이상치들은 실수에 의한 데이터이거나 또는 기타 예외적인 고품질이나 고가 또는 저가의 다이아몬드일지 모른다.

〈표 1〉 분산 분포에 관한 통계적 분석결과의 비교

구 분		$\hat{y} = b0 + b1x + \varepsilon$	$1/\hat{y} = b0 + b1\log x + \varepsilon$
모형 F 값 R2 값 SSE/SSTO 값		$\hat{y} = 3614.71 + 1044.85x + \varepsilon$ F = 26.299 R2 = 0.2133 0.7867	$\hat{y} = (1/0.000254 - 0.0001\log x) + \varepsilon$ F = 61.807 R2 = 0.3892 0.6100
분포특성	·선형특성(Linearity)	비선형, + 방향 증가 분포	선형, - 방향 감소분포
	·잔차(Residual)	증가 부등분산	증가 등분산
	·대칭성 및 이상치 유무(Symmetry and Outlier)	대칭, 1개의 이상치	비대칭, 1개의 이상치

* 주: SSE/SSTO (오차분) = Sum of Squares for Error / Sum of Squares for Total Error

〈표 2〉 이상치(Outlier) 제거 전과 제거 후의 분석결과 비교

구 분		이상치 제거 전	이상치 제거 후	
			이상치 1개 제거 후	이상치 2개 제거 후
모형 F 값 R2 값 SSE/SSTO 값		$\hat{y}=(102/-1.561+0.32\log x) + \varepsilon$ F = 61.857 R2 = 0.3894 0.610	$\hat{y}=(102/-1.549+0.3066\log x) + \varepsilon$ F = 61.760 R2 = 0.3915 0.610	$\hat{y}=(102/-1.5575+0.3099\log x) + \varepsilon$ F = 68.701 R2 = 0.4197 0.580
분포특성	·선형특성(Linearity) ·잔차(Residual) ·대칭성 및 이상치 유무(Symmetry and Outlier)	선형 증가 등분산 대칭, 2개의 이상치	선형 증가 등분산 대칭, 1개의 이상치	선형 증가 등분산 대칭, 이상치 없음

Ⅳ. 최적화 모형의 고찰

위 Ⅱ 및 Ⅲ의 결과로부터 적정 회기분석 모형으로 $\hat{y} = 10^2/(b_0 + b_1\log x) + \varepsilon$ 의 선형식을 선정할 수 있다. 이 경우 등분산성과 이상치를 갖는 비선형 회기모형으로부터 이러한 요인들을 개선시킨 $\hat{y} = 10^2/(-1.5575 + 0.3099\log x) + \varepsilon$ 의 선형 회기방정식을 구하였다.

이제 끝으로 이상치가 없는 위 모형으로부터 2개 이상의 중복 데이터가 있는 경우의 모형 부적격 검정을 행하여보자.

다음 <표 3>은 컴퓨터로부터 얻은 부적합 검정(Lack of Fit Test)을 위한 관련 자료이다. <표3>으로부터 $F^*_{LOF} = 0.83 < F(1-0.001, 17, 19) = 34.172$의 결과는 y에 대한 기대치가 $\{E(y)\} = \beta_0 + \beta_1 x'$로 판단하는 것이 적절함을 나타내는 검정결과를 나타내며 이 식이 부적절한 식이 아님을 입증한다. 따라서 변형모형의 회기식은 선형함수이며 정규분포 및 등분산성을 갖는 적절한 모형임을 추론할 수 있다. 여기서 F^*_{LOF} = MSLF/MSPE의 값이고 n은 총 데이터의 개수, c는 중복 데이터의 개수를 나타낸다.

〈표 3〉 부적합 검정을 위한 컴퓨터 출력

항 목	자승 합(SS)	자유도(df)	평균제곱(MS)	F
회기성(Reg)	SSR=0.0001682	1	MSR=0.0001682	F* = 34.172
에러(Error)	SSE=0.0001772	n-2 =36	MSE=0.0000049	
부적합성(Lack of Fit)	SSLF=0.0000756	c-2 =17	MSLF=0.0000044	F*LOF = 0.83
순수에러(Pure Error)	SSPE=0.0001016	n-c =19	MSPE=0.0000053	
계(Total)		n-1 =37		

주:SSR = Sum of Squares for Regression MSR = Mean Square for Regression
SSE = Sum of Squares for Error MSE = Mean Square for Error
SSLF = Sum of Squares for Lack of Fit MSLF = Mean Square for Lack of Fit
SSPE = Sum of Squares for Pure Error MSPE = Mean Square for Pure Error

V. 결론

이제까지 다이아몬드 구매를 위한 적절한 예측 가격 모형을 구하기 위하여 가격에 영향을 미치는 다이아몬드 데이터의 캐럿(Carat), 색깔(Color), 투명도(Clarity), 품질등급(Cert), 절단상태(Cut) 및 가격(Price)을 고려한 최적화 선형회기 분석모형을 구하였다. 제시된 최적화 모형은 선형분포의 변형모형으로,

$$\hat{y} = b_0 + b_1 \log x + \varepsilon \text{ 즉,}$$

$$\hat{y} = 10^2/(-1.5575 + 0.3099 \log x) + \varepsilon \text{ y} \quad (5.1)$$

임을 통계적 방법을 통하여 검증하였다. 또한 가장 저렴한 가격으로 구매할 수 있는 가격은 가장 낮은 위치의 이상치(Outlier) 데이터로부터 얻을 수 있으며 carat=0.69, price=$1,672인 다이아몬드로 나타났으며 가장 고가의 구매가격은 가장 높은 위치의 이상치 데이터로부터 carat=1.11, price=$14,832인 다이아몬드로 나타났다.

향후 각 변수 간 교호작용(interaction)까지 고려한 선형 다중회기 분석모형(Multiple Regression Analysis Model)에 의한 적절한 가격예측에 관한 연구가 필요하며 적절한 가격예측을 위하여 선형모형의 경우, 정규 분포성, 등분산성 및 오차특성 등을 포함한 연구가 이루어져야 할 것이다.

이러한 연구결과를 바탕으로 무한한 가능성을 지닌 전자상거래 시장에서 예측이 어려운 다이아몬드의 구매예측 가격을 제시할 수 있는 연구방안을 마련할 수 있을 것이다.

참고문헌

[1] O. Lyman, "An Introduction to Statistical Methods and Data Analysis", 3rd ed., PWS-Kent Publishing Co., Boston, 1998.

[2] W. Sanford, "Applied Linear Regression", 2nd ed., John Wiley & Sons, NY, 1997.

[3] A.T. Allen and C.K. Brenda, "SAS/STAT Guide for Personal Computers", Ver. 6 ed., SAS Institute Inc., North Carolina, 1997.

[4] Rainer Sacks, "Rapaport Diamond Report", Newsletter, NY, 1989.

[5] 구상회, 강병구, "인터넷 기반 전자상거래", 고려대학교 출판부, 1999.

Study on the Statistical Optimum Model of Simple Linear Regression to Estimate the Purchasing Price of Diamond

Young-Wook, Lee*

Abstract

The purchasing estimate price of diamond is affected by the factors of carat, color, clarity, certificate, cut and price with the unit of $/carat. The object of this study is to obtain the linear regression model for such purchasing estimate price and to test statistically.

The optimum model is the simple regression model of $\hat{y} = 10^2 / (-1.5575 + 0.3099 \log x) + \varepsilon$ statistically satisfied by the lack of fit test and has the characteristics of normality, constant variance and symmetry.

* Dept. of Computer Science, Semyung University