인공신경망을 이용한 회귀분석 사례 조사

김지현, 이상복
서경대학교 대학원 경영학과

A case study to Regression Analysis using Artificial Neural Network

Jie-hyun Kim, Sang-bok Ree
Dept. of Business Administration, Seokyeong University Graduate School

Abstract

Forecasting have qualitative and quantitative methods. Quantitative one analyze macro-economic factors such as the rate of exchange, oil price, interest rate and also predict the micro-economic factors such as sales and demands.
Applying various statistical methods depends on the type of data. when data has seasonality and trend, Time Series analysis is proper but when it has casual relation, Regression analysis is good for this. Time Series and Regression can be used together. This study investigate artificial neural networks which is predictive technique for casual relation and try to compare the accuracy of forecasting between regression analysis and artificial neural network.

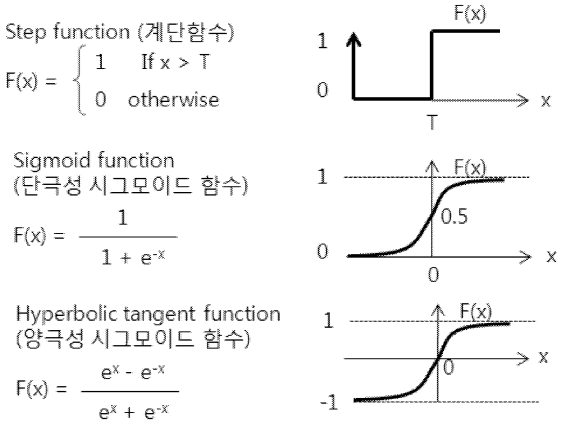
Key Words: regression analysis, qualitative and quantitative methods, artificial neural network

I. 서론

예측은 정책결정과 계획에 영향을 미친다. 한국은행이 경제성장율과 인플레이션을 고려하여 어떻게 금리를 결정할 것인가? 제조업에 근무하고 있는 생산부장이 내년 판매량을 알지 못하고 어떻게 생산 스케쥴을 계획할 것인가? 모든 사람들에게 예측은 필요하다. 예측 방법으로 정성적 기법과 정량적 기법이 있다. 보통 데이터에 기반에 정량적 기법에 관한 연구들이 많이 진행되어 왔다. 특히 예측에 가장 자주 사용되어진 방법은 시계열분석과 회귀분석이다. 시계열분석은 과거 데이터 자신으로부터 미래의 패턴을 발견하는 것이고 회귀분석은 종속변수와 원인변수간의 관계를 파악하여 그 패턴을 찾는 방법이다. 이번 사례조사에서는 예측 방법의 하나인 인공신경망에 대해 살펴 보고 회귀분석 자료에 대해 인공신경망을 적용하여 그 예측력을 비교해 보았다.
인공신경망은 생물의 신경전달 과정을 단순화하고 이를 수학적으로 해석한 모델로써, 복잡하게 얽혀있는 뉴런(neuron)을 통과시켜가면서 뉴런끼리의 연결강도를 조절하는 일종의 학습(training)과정을 통하여 새로운 문제를 분석

한다. 이러한 과정은 사람이 학습하고 기억하는 과정과 비슷하며 이를 통해 추론, 분류 그리고 예측을 수행하는 것이다.
예를 들어, 지난 3년 기간의 주택담보대출에 대한 상환이행에 대한 정보가 있다고 해 보자. 이 정보에는 직업, 소득, 거주 지역, 학력, 대출액 등등에 대한 입력 정보와 만기일에 상환되었는지, 지연되었는지 혹은 불이행 되었는지에 대한 과거 종속 정보들이 있을 것이다. 이런 정보를 학습하여 새로운 신청자가 왔을 때 이 사람의 입력 정보만을 가지고 사전에 상환 가능성을 예측해 볼 수 있다.
지금까지 이러한 문제들은 과거의 데이터를 이용하여 확률 또는 통계를 통하여 문제를 해결하는 방법이 일반적으로 사용되어 왔다. 그러나 인공신경망을 이용하여 접근하면 훨씬 더 간단하고 쉽고, 비교적 정확히 답을 구할 수 있다. 무엇보다도 통계에 대한 전공자가 아니라도 상용화된 소프트웨어를 이용한다면 간단히 예측을 실시할 수 있다는 장점이 있다.
신경망은 인간두뇌가 반복적인 경험을 통한 학습이 내적 지식으로 습득하는 것과 마찬가지로 학습 능력이 있고, 외부의 자극에 대한 신경조직의 전달체계가 병렬 처리하는 단계를 가지고 있다. 또한 경험에서의 편견이나 왜곡된 자료에 대해서도 기존의 더 큰 자극에 대해서 반응하므로 결함허용시스템을 내적으로 지니고 있는 인공지능의 한 분야이다.
인공신경망은 은닉마디라고 불리는 독특한 구성요소에 의해서 일반적인 통계모형과 구별된다. 은닉마디는 인간의 신경세포를 모형화한 것으로 각 은닉마디는 입력변수들의 결합을 수신하여 활성함수라는 비선형 함수를 통해 목표변수에 전달한다. 이 때 결합에 사용되는 계수(coefficient)를 연결강도(synaptic weights)라 하며 활성함수는 입력값을 변환하고 이를 입력으로 사용하는 다른 마디로 출력하는 역할을 한다. 또한 인공신경망은 학습규칙이라는 반복적 알고리즘을 통하여 주어진 문제를 해결해 나간다.

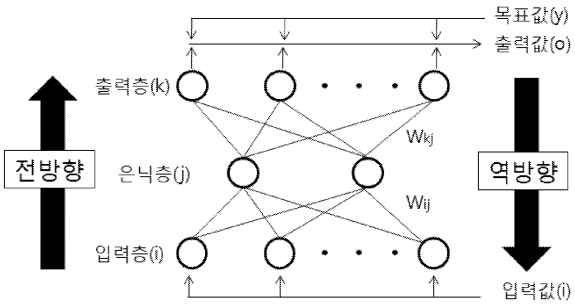
인공신경망 모형은 적용되는 신경망의 종류, 학습규칙 그리고 적용되는 활성함수의 종류에 따라 다양하게 분류될 수 있다. 따라서 인공신경망 알고리즘을 적용하기 위해서는 특정 문제해결에 적합한 알고리즘과 학습규칙, 활성함수의 선택이 필요하다.



<그림 1> 활성함수의 종류

다층 전방향 퍼셉트론(Multi-Layer Feedforward Networks)은 역전파 알고리즘으로서 복잡한 기능을 근사화하는 기능을 가지고 있다. 그래서 독립변수들과 종속변수간의 복잡한 관계를 모형화 해준다. 이 학습 방법은 지도학습(supervised learning)이다. 즉, 학습을 하기 위해서는 입력 데이터(i)와 원하는 출력(o) 데이터가 있어야 한다. 간단히 학습의 개념을 살펴보면 먼저, 입력이 신경망의 가중치(weights)와 곱하고 더하는 과정을 몇 번 반복하면 입력의 결과값이 출력(y)으로 나온다. 이때 출력(y)은 학습 데이터에서 주어진 원하

는 목표(o)와 다르다. 따라서 인공신경망의 학습에서 (y-o)만큼의 오차(e)가 발생하며, 오차에 비례하여 출력층의 가중치를 갱신하고, 은닉층의 가중치를 갱신한다. 가중치를 갱신하는 방향이 신경망의 처리 방향과는 반대 방행이다. 즉 신경만의 처리는 입력층 -> 은닉층 -> 출력층의 방향으로 진행되며, 가중치 갱신의 학습방향은 출력층 -> 은닉층으로 진행된다. 이런 이유로 역전파 알고리즘이라고 한다.



<그림 2> 역전파 알고리즘

역전파 알고리즘의 단계를 간단히 정리하면 아래와 같다.

1. 신경망에 입력 데이터를 입력 노드에 적용하고, 입력에 다른 출력을 계산한다.
2. 입력에 따른 출력과 원하는 출력간의 오차를 계산한다.
3. 오차를 줄이기 위해 가중치의 증감 여부를 결정한다.
4. 각각의 가충치를 얼마나 변화시킬 것인가를 결정한다.
5. 4 단계에서 결정된 값으로 가중치를 갱신(변화)한다.
6. 모든 학습 데이터에 대해 오차가 적정 수준으로 감소하기까지 1단계에서 5단계를 반복한다.

하지만 역전파 알고리즘은 다음과 같음 몇 가지 문제점을 지니고 있다.

첫째, 역전파 알고리즘이 지역적 최소점(local minimum)에 빠질 우려가 있다. 역전파 알고리즘의 기본적인 원리는 최급하강법(gradient descent method)이다. 즉 최급하강법에 의한 학습은 전역적 최소점(global minimun)을 목표로 해서 가중치를 조정하는 것이 아니라 현재의 위치에서 경사면을 따라 내려가는 것이기 때문에 오차가 0이 아닌 지역적 최소점에서 더 이상 학습결과를 개선하지 못하고 정체될 수 있다. 두 번째 문제점은 전방향 신경망과 관련된 문제로서, 이는 이 방법이 기본적으로 신호과 입력에서 출력까지 한 방향으로만 진행되는 전방향 방식이라는 것이다. 이러한 전방향 신경망은 고차원의 시스템을 구성하는데는 한계가 있는 것으로 알려져 있다. 이외에도 역전파 알고리즘은 학습이 완료되기까지 많은 회수와 반복학습과 응용분야에 따라 상이하게 학습모수를 조정하는 것이 필요하고, 추가 학습시 전체적인 재학습이 요구된다. 또한 학습의 완료시점을 예측할 수 없다는 문제점을 가지고 있다.

II. 본론

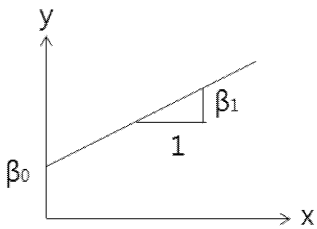
1. 선형회귀 자료에 대한 신경망 적용

회귀분석이란 변수들 중에서 주된 관심이 되는 변수 하나를 종속변수로 설정하고 너머지 변수들을 독립변수로 하는 선형모형의 유의성을 규명하는 통계적 기법이다. 상관분석이 둘 이상의 변수들이 어느 정도 상관관계를 갖고 있는가를 파악하는데 목적이 있는 반면 회귀분석은 독립변수와 종속변수의 선형 관계식을 통해 종속변수의 값을 예측하는데 주된 목적이 있다. 즉, 회귀분석은 변수들 간의 관련성을 규명하기 위하여 독립변수(x)와 종속변수(y)간의 선형함수관계를 가정하고 데이터로부터 이 함수

를 추정하여 종속변수와 관련된 통계적 추론을 하는 통계적 분석방법인 것이다.
회귀모형의 종류는 종속변수의 형태와 독립변수의 수에 따라 다음과 같이 분류될 수 있다.

■ 독립변수의 수에 의한 분류
단순 선형회귀 : 독립변수(x)가 한 개
다중 선형회귀 : 독립변수(x)가 두 개 이상

■ 반응변수의 종유에 의한 분류
선형 회귀 : 반응변수(y)가 연속형
로지스틱 회귀 : 반응변수(y)가 범주형



<그림 3> 단순 선형회귀 그래프

1) 선형 회귀자료에 대한 신경망 분석 실시
표1은 5개의 독립변수(x)에 대한 1개의 종속변수(y)의 타이어 발열량 데이터이다. 데이터셋의 맨 마지막 행을 예측하고자 하는 데이터로 가정하고 분석을 실시해 보았다.
먼저 통계분석 소프트웨어인 Minitab을 통하여 다중선형회귀 분석을 실시해 보았다.
5개의 독립변수 중에 종속변수에 영향을 미치는 유의한 변수를 한 번에 찾기 위한 방법으로서 Minitab의 단계적 회귀분석을 실시하였고 하중(x1)과 속도(x2)만이 유의한 영향을 미치고 있음을 파악하여 이 2개의 독립변수에 대해서 다시 한번 잔차 분석과 분산 팽창 인수 분석을 동시에 수행하여 최종적으로 그림4와과 같은 결과을 얻었다.

<표 1> 선형 회귀자료 데이터

하중(x1)	속도(x2)	shoulder 두께(x3)	지표온도(x4)	주행시간(x5)	발열량(y)
70	68	36.5	36	5	91
72	72	36	36	6	89
75	93	37	37	6	105
71	97	36.3	37	6	106
78	119	36.5	39	4	113
76	113	36	39	5	114
89	79	36.5	38	5	117
92	75	36.3	38	6	115
91	76	36.6	39	5	125
96	92	36.6	39	6	126
97	113	37	38	6	140
95	110	35.6	38	6	141
106	68	35.3	38	7	140
109	74	36.8	35	7	142
112	92	35.3	38	5	150
115	95	35.3	38	6	149
117	114	37.1	38	4	168
114	112	35.6	37	5	166

회귀 분석: 발열량(y) 대 하중(x1), 속도(x2)

회귀 방정식
발열량(y) = - 21.7 + 1.26 하중(x1) + 0.347 속도(x2)

예측 변수	계수	계수 SE	T	P	VIF
상수	-21.662	8.121	-2.67	0.018	
하중(x1)	1.25626	0.06716	18.70	0.000	1.004
속도(x2)	0.34725	0.06110	5.68	0.000	1.004

S = 4.32609 R-제곱 = 96.6% R-제곱(수정) = 96.1%
<그림 4> 다중선형 회귀 분석 결과 자료

그런 후 새로운 관측치인 하중 114와 속도 112에 대한 예측치를 회귀식에 입력한 결과 적합지가 160.44를 얻게 되었다. 원 데이터의 실제 관측치는 166이었고 예측치(적합치)는 160.44으로써 오차가 5.56가 발생하고 95% 신뢰구간에서 115.79에서 165.10의 값을 가질 수 있음을 파악하였다.

새로운 관측치에 대한 예측치

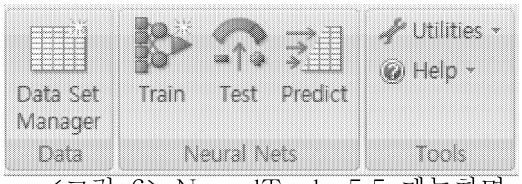
새로운 관측치	적합치	SE 적합치	95% CI	95% PI
1	160.44	2.17	(155.79, 165.10)	(150.06, 170.82)

새로운 관측치에 대한 예측 변수의 값

새로운 관측치	하중(x1)	속도(x2)
1	114	112

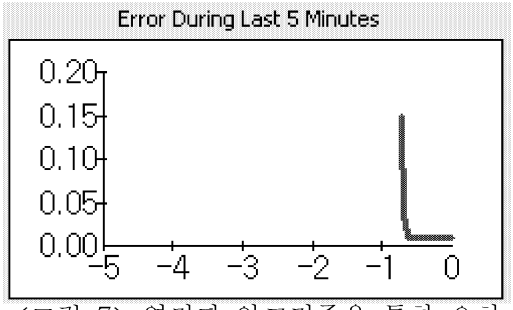
<그림 5> 새로운 관측치에 대한 적합치

다음으로 동일한 자료에 대해 신경망 모형을 적용시켜 보았다. 신경망 모형 분석에 대표적인 소프트웨어인 NeuralTools 5.5를 사용하였으며 본 소프트웨어는 데이터 셋 지정, 학습, 테스트 그리고 예측 단계로 분석이 수행된다.



<그림 6> NeuralTools 5.5 메뉴화면

서론에서 신경망 모형의 역전파 알고리즘을 설명한바와 같이 출력값과 목표값이 0에 가까운 가중치를 찾는 과정을 소프트웨어에서 자동으로 실시하게 된다.



<그림 7> 역전파 알고리즘을 통한 오차 최소화

오차가 최소화가 되었을 경우 자동 종료되며 우리가 알고자 했던 하중 114, 속도 112,shoulder 두께 35.6, 지표온도 37, 주행속도 5 인 경우에 대한 신경망은 160.51을 예측하여 거의 오차가 없었다.

하중(x1)	속도(x2)	shoulder 두께(x3)	지표온도(x4)	주행시간(x5)	발열량(y)
114	112	35.6	37	5	

Prediction Report: "

Tag Used	Prediction
predict	160.51

<그림 8> 단순회귀 자료에 대한 신경망 분석결과

2) 선형회귀와 신경망에 대한 비교

선형회귀와 신경망 분석을 상용 소프트웨어를 활용하여 분석해 보았다. Minitab에서 단계적 회귀분석 기능을 이용하여 유의한 설명변수를 찾은 후 유의한 모형을 얻을 수 있다. 신경망 모형은 독립변수들과 종속변수들에 대한 데이터셋을 지정하여 자동으로 학습을 실시하여 예측할 수 있었다. 예측에 대한 비교결과 신경망이 단순선형 회귀보다는 좋은 예측력을 보여주고 있다. 기타 선행 논문들을 살표보면 신경망이 회귀분석보다 예측 정확도가 높은 결과를 제시한 사례를 많이 찾아볼 수 있다. 하지만 이것은 경우에 따라 다를 수 있기 때문에 단언할 수 없으며 이에 대한 별도의 연구가 필요할 것이다.

2. 로지스틱 자료에 대한 신경망 적용

선형 회귀분석은 종속변수가 연속형 데이터인 경우 적용해 보았다. 종속변수가 이진값 예를 들어, 실패/성공, 정품/불량 등과 같은 경우 사용되는 방법이 로지스틱 회귀분석이다. 로지스틱 회귀분석의 설명변수는 측정형과 분류형이 가능하지만 지시 변수가 너무 많으면 모형이 복잡해지고 해석이 복잡해진다.

로지스틱 회귀분석에서는 종속변수 값은 0,1 (예: 성공, 실패)로 입력된다. 로지스틱 회귀분

석은 이진형 반응변수 뿐 아니라 반응변수가 순서형(ordinal)인 경우 사용할 수 있다. 예를 들어 종속 변수가 고객의 신용도인 경우 상, 중,하로 분류되어 있는 경우도 사용할 수 있다.

2-1 로지스틱 이항 회귀자료에 대한 신경망 분석 실시

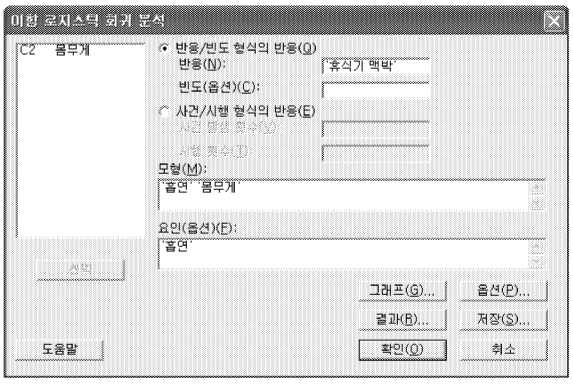
표2는 총 91개의 데이터를 제시하고 있으며 2개의 독립변수(x)에 대한 1개의 종속변수(y)인 휴식기 맥박의 데이터이다. 데이터 셋의 맨 마지막 2행을 예측하고자 하는 데이터로 가정하고 분석을 실시해 보았다.

<표 2> 휴식기 맥박 데이터

흡연	몸무게	휴식기 맥박
아니오	64	낮음
아니오	66	낮음
예	73	낮음
예	86	낮음
아니오	70	낮음
(중략)		
예	49	낮음
아니오	43	높음
예	57	높음
아니오	60	낮음
아니오	50	낮음
아니오	49	낮음

변수를 살펴보면 독립 변수에 흡연(예, 아니오)과 종속변수에 휴식기 맥막(높음, 낮음)이 범주형 자료이므로 이항 로지스틱 회귀분석을 먼저 Minitab으로 실시해 보았다.

그림9는 Minitab 메뉴에서 이항 로지스틱 회귀분석의 대화상자와 분석결과이다.



<그림 9> Minitab 이항 로지스틱 회귀분석

로지스틱 회귀 분석 표

예측 변수	계수	계수 SE	Z	P	승산비	95% CI 하한	95% CI 상한
상수	2.61881	1.78649	1.47	0.143			
몸무게	-0.0658868	0.0287992	-2.29	0.022	0.94	0.88	0.99
흡연							
예	1.28479	0.568622	2.26	0.024	3.61	1.19	11.02

로그 우도 = -44.042
모든 기울기가 0인지 검정: G = 9.170, DF = 2, P-값 = 0.010

적합도 검정

방법	카이-제곱	DF	P
Pearson	38.6685	46	0.770
이탈도	48.2796	46	0.381
Hosmer-Lemeshow	4.8700	8	0.771

화면

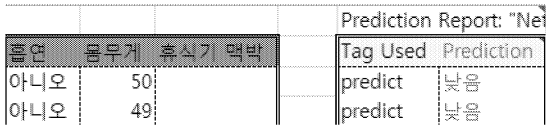
<그림 10> 이항 로지스틱 회귀분석 결과 자료

분석 결과인 그림10을 보면 흡연(예)와 몸무게가 p-value 0.05이하로서 유의함을 얻었다. 그래서 회귀식은 y = 2.61881 +1.28479*흡연(예) - 0.0658868*몸무게 이다. 새로운 관측치에 대한 예측을 위해서 이 회귀식 y를 지수변환 해야 하며 변환식은 exp(y)/[1+exp(y)] 이다.

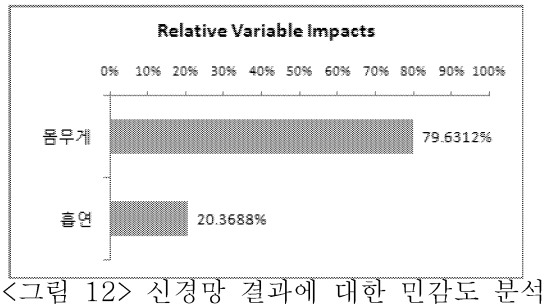
변환 결과가 0.5 이상이면 휴식기 맥박(높음), 0.5 미만이며 휴식기 맥박(낮음)이다. 따라서 새로운 관측치 흡연(아니오), 몸무게(50)은 y=-3.29434이고 변환값은 0.035766으로서 휴식기 맥박 낮음을 예측할 수 있다.

마찬가지로 흡연(아니오), 몸무게(49)도 휴식기 맥박 낮음을 얻었다. 따라서 원 데이터 자료와

비교해 봤을 때 동일한 결과를 얻게 된 것을 확인하였다.
다음으로 신경망 모형을 통하여 분석해 보자. 마찬가지로 원 데이터에 대한 학습과 훈련을 실시하였고 그 결과는 아래와 같았다.
그림11에서 보듯이 흡연과 몸무게의 두 경우에 대해 휴식기 맥박이 낮음을 예측할 수 있었다.
더불어 민감도 분석을 실시해 본 결과 몸무게가 휴식기 맥박에 중대한 영향을 미치고 있음을 알 수 있다.



<그림 11> 이항 로지스틱 자료에 대한 신경망 분석 결과



<그림 12> 신경망 결과에 대한 민감도 분석

2-2 로지스틱회귀와 신경망에 대한 비교
로지스틱회귀와 신경망 예측에 대한 결과는 똑같았다. 하지만 분석 방법상에서 통계 전공자가 아닌 경우 신경망 분석이 보다 쉽게 수행할 수 있다는 장점이 있었다.

3. 결론

예측 방법에 있어서 대표적으로 사용되고 있는 회귀분석 방법은 독립변수와 종속변수간의 관계를 파악하여 그 패턴을 파악하고 새로운 관측치에 대한 예측을 수행하게 된다. 이번 연구에서는 인과분석(casual analysis) 방법 중 하나인 인공신경망을 적용하여 동일한 자료에 대해 회귀분석 결과와 비교해 보았다. 본 연구에서 사용한 사례만을 가지고 비교한 결과 인공신경망이 선형회귀보다는 예측력이 높음을 파악할 수 있었고 로지스틱회귀와는 동일한 결과를 얻게 되었다. 하지만 연구의 한계로서 다양한 사례를 적용하지 못했기 때문에 예측력에 대한 우수성을 단정하기 어려우며 보다 많은 사례에 적용해 보아 비교 및 평가가 필요하다. 하지만 방법상에 있어서 인공신경망이 분석 방법 상에서 간단하여 실무에 쉽게 사용할 수 장점이 있었다.

참고문헌

[1] 이동수(2008), “인공신경망을 활용한 개별공시시자 산정방법 개선에 관한 연구” 영남대학교 대학원 박사학위논문
[2] 양호원(2008), “신경망과 의사결정트리를 이용한 Stream Data 예측 시스템 설계 및 구현”, 조선대학교 대학원 박사학위논문
[3] 배장한(2009), “영상분할 방법 기반의 인공신경망을 적용한 카메라의 렌즈왜곡 보정”, 성균관대학교 대학원 석사학위논문
[4] 류인환(2006), “인공신경망에 의한 시계열 자료의 수요예측”, 연세대학교 대학원 석사학위논문
[5] 유성모, 박현주 공저 (2006), 「Minitab으로 배우는 기초통계」, 이레테크
[6] Palisade Corporation (2005), 「Guide to NeuralTools Manual」