論文
10-5-6

MS Excel 함수들을 이용한 회귀 분석 모형 추정 및 관계 분석 검정을 위한 매크로 개발
(지하철 전기요금 자료 회귀분석에 응용)
Development of MS Excel Macros to estimate regression models and test hypotheses of relationships between variables (Application to regression analysis of subway electric charges data)

김숙영(Kim Sook Young)[1)]

요약

변수들간의 관계 모형을 설정하고 관계성 유무를 분석하는 회귀 분석은 거의 모든 조사 연구 및 실험연구들에서 필수적인 통계 분석 방법이다. 자료는 독립변수와 종속변수로 구성되므로 쌍으로 취급되며 모든 통계량 계산은 행렬 연산에 의하여 수행된다. 변수들 관계를 가장 잘 설명하는 모형 설정에 따라 회귀분석 결과의 정확성이 평가되므로 자료 수치들을 XY 평면상에서 점을 찍어 가장 적합한 함수 모형을 선택해야 한다.
MS 엑셀의 그래픽 및 행렬 연산 기능의 메뉴들을 사용하면 수집된 자료에 가장 적합한 모형을 설정하고 필요한 모든 가설검정 작업을 쉽게 수행할 수 있다.
본 연구에서는 회귀 분석의 모형 설정 및 가설검정 결과들을 산출하는 엑셀 함수를 이용한 매크로를 개발하였다. 본 연구에서 개발한 회귀분석 매크로를 한 개의 종속변수와 3개의 독립 변수를 가진 지하철 전기요금 자료 분석에 적용하여 얻은 결과와 엑셀에 내장된 통계 회귀분석 메뉴를 적용한 결과를 비교한다.

Abstract

Regression analysis to estimate the fitted models and test hypotheses are basic statistical tools for survey data as well as experimental data. Data is collected as pairs of independent and dependent variables, and statistics are computed using matrix calculation. To estimate a best fitted model is a key to maximize reliability of regression analysis. To fit a regression model, plot data on XY axis and select the most fitted models. Researchers estimate the best model and test hypothesis with MS Excel's graph menu and matrix computation functions.
In this study, I develop macros to estimate the fitted regression model and test hypotheses of relationship between variables. Subway electric charges data with one dependent variable and three independent variables are tested using developed macros, and compared with the results using built-in Excel of regression analysis.

Key words: Regression analysis, Excel functions, Statistics, Matrix computation

논문접수 : 2009. 12. 01.
접수완료 : 2009. 12. 27.

1) 정회원 : 안산공과대학 컴퓨터정보과

1. 서론

IT 산업 발달로 인간의 사고도 디지털화됨에따라 기업체와 정부의 의사결정 및 정책수립에서도 확률적 접근이 필요하다. 특히 학문분야에서는 실험에 근거한 연구방법을 도입하는 농업, 생명과학, 환경과학,자연과학 및 산업과학은 물론 어떤 특정한 현상(주제, 사실)을 조사하기 위하여 연구실 밖으로 나가 실제로 존재하는 것을 조사하는 인구 및 주택센서스 여론조사 교통량 조사 등 문학적면을 다루는 인문 사회과학에도 자료를 양적 취급하는 확률적 연구 방법을 적용하고 있다.[1]

통계학 (Statistics)는 확률을 뜻하는 라틴어의 statisticus(확률) 또는 statisticum(상태),특히 이탈리어의 statista(나라, 정치가) 등에서 영향을 받아 국가의 인력, 재력 등 국가적 자료를 비교 검토하는 학문을 의미하게 되었다.

현재는 매우 다양한 분야의 연구에서 주어진 문제에 대하여 적절한 정보를 수집하고 응용수학 기법을 이용하여 수치상의 성질, 규칙성 또는 불규칙성을 찾아낸다.

자료의 형태에는 양적인 성격의 수치로 표현가능한 수량자료로 일정한 범위내에서 값을 가질 수 있는 등간척도 자료 및 절대 양 뿐 아니라 비율형태 값을 가질 수 있는 비율 척도 자료가 있다.

질적인 성격으로 수치 측정이 불가능한 분류자료는 단순한 번호로, 명의척도로 순서의 의미는 없다. 자료는 행과 열로 셀들이 구성되어 상대 또는 절대 참조 방식에 의하여 데이터와 수식을 담을 수 있는 셀을 유지하고 있는 스프레드시트인 엑셀은 직관적인 인터페이스 기능 및 계산 기능 및 그래픽도구들로 가장 인기있는 프로그램이 되었다.

스프레드 시트의 외양(글꼴, 글자속성 및 외양)을 사용자가 지정할 수 있는 최초의

회귀분석은 시간에 따라 변하는 데이터나 스프레드시트이고 매번 또는 사용자 명령이 있을때만 재계산을 했던 과거의 스프레드시트 또는 프로그램에 대신 연관된 셀들에 대하여 재계산을 수행하는 기능으로 발전되었다.

1993년부터 엑셀은 비주얼베이직에 기반하여 엑셀의 자동화와 워크시트에서의 사용자 지정 함수(UDF, User defined function)사용을 가능하게 한 Visual Basic for Application(VBA) 언어를 포함한다.

이후 버전에서 통합 개발환경 (IDE, Integrated development environment)를 제공하는 VBA는 이 프로그램에 아주 큰 장점이 되었다.

사용자의 움직임을 재수행할 수 있는 매크로 녹음 기능은 일반적으로 작업의 자동화를 가능하게 하였다. VBA를 사용자와 대화할 수 있는 폼 (form) 과 워크시트 내 컨트롤 생성을 가능하게 한다. 이 언어는 ActiveX(com) 과 동적 연결 라이브러리 사용을 지원하며 이후 버전에서는 기본적인 객체지향 프로그래밍 기술을 사용할 수 있도록 클래스 모듈을 지원한다.[2]

현실에서의 많은 형상들을 인과 관계에 따라 관계성 유무를 확률적으로 검정하고 미래의 현상을 예측하는 분석 방법인 회귀분석이 가장 많이 사용된다.

회귀분석은 영국의 유전학자 프란시스 갈톤 (Francis Galton)이 부모의 키와 자식의 키 사이에 연관 관계가 있을 것이라는 추론하에 연구하면서 키가 커지거나 작아지는 것보다 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세우고 분석하는 방법을 연구에서부터 시작되었다. 본격적인 연구는 칼 피어슨 (Karl Pearson)의 아버지와 아들의 키 사이에 함수 관계를 도출하여 수학적 관계를 성립하였다.

관심있는 통계적 추론으로 활용된다.

어떤 영향, 가설적 실험, 인과 관계의 모델링등의 통계적 예측에 이용될 수 있다. 그러나 회귀분석은 반드시 수학적 모형의 설정이 필요하므로 가정이 맞는 지 아닌 지 적절하게 밝혀지지 않은 채로 이용되어 결과가 오용되는 경우가 많다. 회귀분석 통계량 수치적 계산은 행렬연산에 의하여 수행되므로 다양한 함수들을 가

지고 있는
엑셀에서 필요한 계산 매크로를 개발하여 분석에 이용할 수 있다.[3]
통계 패키지 대신 엑셀을 이용하여 회귀분석을 수행하면 엑셀의 편리한 그래픽 기능으로 회귀 모형 설정이 용이하고, 필요한 결과만을 얻을 수 있으며 결과 해설이 쉽다는 이점을 가지고 있다.
특히 조사연구 자료들은 변수들간의 관계를 가장 잘 설명하는 모형을 찾기 어려우므로 자료값들을 XY 그래프상에 점을 찍어 함수를 선택하는 작업을 여러번 시행하여야한다. 이 작업을 가장 쉽게 수행할 수 있는 소프트웨어는 바로 엑셀이다.

본 연구에서는 신뢰성있는 회귀 분석 결과를 쉽게 얻기 위하여 엑셀 함수들의 사용법 및 얻을 수 있는 결과의 범위를 예제와 함께 설명한다.

2. 조사방법

2.1. 이론적 배경

2-1-1. 회귀 분석의 정의

회귀 분석은 원인과 결과의 독립변수(X) 와 종속변수(Y)들간의 관련성을 확률적으로 추정하기 어떤 수학적 모형을 가정하고 이 모형을 측정된 데이터로부터 추정하는 통계적 분석방법이다. 회귀분석을 통해 추정된 모형으로부터 필요한 예측을 하거나 입력 변수 개수에 따라 한 개인 경우는 단순회귀분석, 두 개 이상인 경우는 중회귀분석으로 분류되며, 입력 변수와 종속변수 관계에 따른 분류는 직선 관계로 설정하여 분석되는 경우는 선형회귀, 직선 관계로 설정되어 분석하는 경우는 비선형으로 분류된다.
변수들간의 변동이 수치적 계산의 요소이다. 자료간의 총 변동을 회귀 모형에 의한 변동과 단순한 오차에 의한 변동으로 분리하여 적합한 모형을 설정하고 그 모형에 의하여 변수들간의 관계를 검정하고 예측한다.
추정된 회귀 모형이 변수의 관계를 어느 정도 설명할 수 있는가 하는 정도를 나타내는
0 과 1 사이의 값을 가지는 결정계수는 회귀 분석에서 반드시 필요한 요인이다.
예로 결정계수가 0.8이면 전체 산포도 중 회귀 모형이 80%가 설명되고 나머지 20%는
다른 원인에 의한 것으로 해설된다.

2-1-2. 단순회귀분석

한 개의 독립변수(X)와 종속변수(Y)관계를 분석하므로 모형은 $Y=a+bX+\varepsilon$ 이며, 각 계수의 추정식은

$$b=\frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^2-(\sum x)^2}$$

$$a=\frac{\sum y-b\sum x}{n}$$ 로 유도된다.

변수간의 관계 유무 검정을 위한 가설은

H_o : $\beta = 0$

H_A : $\beta \neq 0$ 이다.

검정통계량 계산식은

$$(y_i-\bar{y})=(\hat{y}-\bar{y})+(y_i-\hat{y})$$

총 변이 = 설명편차 + 오차

의 이론에 의하여

각 항들을 제곱하여 합하면 총 제곱합 = 설명될 수 있는 제곱합 + 오차 제곱합 이 유도되며 회귀분석 용어를 사용하면

SST = SSR + SSE로 표시된다.
총제곱합 = 회귀모형에의한 제곱합 + 오차제곱합

회귀 모형의 가장 중요한 평가 기준은

$$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y_i-\hat{y})^2} = SSR/SST$$

에 의하여 산출되는 결정계수이다.

회귀분석이 두 변수관계를 설명하는데

적합한 지 여부는 계산된 제곱합들로 표 와 같은 분산분석표를 작성하여 가설검정할 수 있다.

<표 1> 단순회귀분석 분산분석표

변 이	제곱합	자유도	평균제곱합	비
선형 회귀	SSR	1	SSR/1	MSR/MSE
오차	SSE	n-2	SSE/(n-2)	
총 합	SST	n-1		

독립변수와 종속변수의 관계유무는 회귀계수(β)의 가설검정을 위한 검정통계량은

$$t=\frac{b_1}{s(b_1)} \sim t(n-1)$$

$$s_b^2=\frac{n-1}{n-2}(s_y^2-b^2s_x^2)$$

(s_y^2: y 값의 분산, s_x^2 : x 값의 분산) 이다. 변수들간의 관계유무 (β=0) 뿐 아니라 양의 관계인가 음의 관계인가 까지 검정할 수 있다.

2-1-3. 다중 회귀 분석

독립 변수가 두 개 이상이고 종속변수가 한 개인 회귀분석을 다중 회귀 분석이라하며
회귀 모형은

$$Y_i=\beta_o+\beta_1X_1+\beta_2X_2+.....\beta_{k-1}X_{k-1}+\epsilon$$

로 추정된다.
회귀 분석에서 검정통계량은 행렬을 이용하여 간편하게 계산된다.

$$Y=\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ Y_k \end{bmatrix} \qquad X=\begin{bmatrix} 1 & X_{11} & X_{21} & ..... & X_{k1} \\ 1 & X_{12} & X_{22} & ..... & X_{k2} \\ & & ... & & 1 \\ X_{1n} & X_{2n} & ... & & X_{kn} \end{bmatrix}$$

$$\beta=\begin{bmatrix} \beta_o \\ \beta_1 \\ ... \\ \beta_{k-1} \end{bmatrix} \qquad \epsilon=\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_k \end{bmatrix}$$

$$Y = X\beta + \varepsilon \qquad E(Y) = X\beta$$

$$X'Xb=X'Y$$

$$b=(X'X)^{-1}X'Y$$

$$\hat{Y}=HY$$

$$H=X(X'X)^{-1}X'$$

단순회귀모형의 경우와 같이 중회귀분석 모형이 적합한 가설검정은 단순회귀모형의 경우와 같이 분산분석표에 의하여 계산된다

<표 2> 중회귀분석 분산분석표

요인	SS	df	MS	F
회귀	SSR	k	MSR=SSR/k	MSR/MSE
오차	SSE	n-k-1	MSE=SSE/(n-k-1)	
총합	SST	n-1		

독립변수가 두개 이상인 다중회귀분석에서는 각 독립변수와 종속변수 관계 유의성에관한 가설검정이 필요하다.

<표 3> 회귀모형 계수들의 가설검정

요인	계수	표준오차 (se)	t 값	p 값
절편	B(0)	s2b(0,0)	계수/se	=TDIST(T값,n-k-1,2)
X1	B(1)	s2b(1,1)	계수/se	0.000675676
X2	B(2)	s2b(2,2)	계수/se	0.015761754
X3	B(3)	s2b(3,3)	계수/se	0.084312621
Rsquare	SSR/SSE		Adjusted Rsquare	=1-SSE/SSTO* ((n-1)/(n-k-1))

다중 회귀 분석에서는 자료가 행렬로 입력되므로 검정통계량 계산도 행렬연산에 의하여 수행되어야한다.

$$SSTO = Y'Y - (\frac{1}{n})Y'JY = Y'I - [(\frac{1}{n})J]Y$$

$$SSE = Y'(I-H)Y$$

$$SSR = Y'[H - (\frac{1}{n})J]Y$$

$$s^2(b) = MSE(X'X)^{-1}$$

$$R_a^2 = 1 - (\frac{n-1}{n-p-1})\frac{SSE}{SSTO} \quad [4]$$

2-2. 매크로 개발

k 개의 독립 변수, 1개의 종속변수를 가진 n 쌍의 자료로 중회귀 분석 시행 시 입력 자료를 표 4 과 같이 정의한다.

<표 4> 입력 자료 행렬 정의

행렬 이름	행렬크기	성격
X (독립변수)	행: n 열 : k+1	첫 번째열은 모두 1의 값
Y (종속변수)	행: n 열: 1	
J(계산에필요)	행: n 열: n	모든 원소들이 1의 값을 가지고 있는 행렬

중회귀 회귀 모형의 분산분석 및 계수 t검정에 필요한 제곱합등의 모든 통계량 계산을 위한 매크로는 표 5에 있다.

<표5> 통계량 계산을 위한 매크로 정의

수식	매크로 명령	크기
① = (XTX)	mmultiply(TRANSPOSE(x), x)	행: k 열: k
②= (XTX)-1	minverse(A)	행: k, 열: k
③ = B행렬 ((k+1)x1)	=MMULT(MMULT(②,XT),Y)	행: k+1 열: 1
SSTO	=MMULT(YT,Y)-MMULT(MMULT (YT,J),Y)/n	숫자
SSE	=MMULT(YT,Y)-MMULT(MMULT (BT,XT),Y)	숫자
SSR	SSTO-SSE	
④= s2(b)	SSE/(n-k-1)x②	행: k+1 열: k+1

표2와 표3 에서의중회귀 모형 분산분석표와 계수 t 검정에 필요한 제곱합들은 표5의 매크로들로 모두 계산될 수 있다. 표 2 의 중회귀 분산분석표의 유의성 검정 p 값은 =FDIST(F값,k,n-k-1)에 의하여 얻어질 수 있다. [2,5]

2-3. 자료 분석에 응용

지하철 주행거리, 전동차 전력, 부대설비 전력이 전력 요금에 미치는 영향을 본 연구에서 개발한 매크로를 적용하여 중회귀 분석 모형에 의하여 분석하고 엑셀의 회귀분석 메뉴를 적용한 결과와 비교한다.

3. 결과 및 요약

3-1. 매크로를 이용한 중회귀분석

전력요금을 종속변수, 지하철 주행거리, 전동차 전력, 부대설비 전력을 독립변수로

<표 6> 매크로를 이용한 전력요금 중회귀 모형 분산분석표

분산분석표					
요인	자유도	제곱합	제곱평균	F값	p값
회귀	3	4,613,371,495,921,280	1,537,790,498,640,430	25.54094461	0.000188981
잔차	8	481,670,673,371,328	60,208,834,171,416		
합계	11	5,095,042,169,292,610			

<표 7> 매크로를 이용한 각 독립변수와 종속변수의 t검정

요인	계수	표준오차	t 값	p 값
절편	107596226.6	79946410.69	1.345854375	0.215243171
X1	4.366014641	0.814181068	5.362461511	0.000675676
X2	2.262311547	0.741140918	3.052471523	0.015761754
X3	-7.724950205	3.920821067	-1.970237885	0.084312621
Rsquare	0.905462868		Adjusted Rsquare	0.870011444

<표 8> 엑셀 회귀분석 메뉴 적용 결과

회귀분석 통계량	
다중 상관계수	0.951558
결정계수	0.905463
조정된 결정계수	0.870011
표준 오차	7759435
관측수	12

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	3	4.61E+15	1.54E+15	25.54094	0.000189
잔차	8	4.82E+14	6.02E+13		
계	11	5.1E+15			

	계수	표준 오차	t 통계량	P-값
Y 절편	1.08E+08	79946411	1.345854	0.215243
X 1	4.366015	0.814181	5.362462	0.000676
X 2	2.262312	0.741141	3.052472	0.015762
X 3	-7.72495	3.920821	-1.97024	0.084313

중회귀 분석을 수행한 결과는 표6, 7에 있다. F=25.54 p=0.0001로 세 개의 독립변수들이 전력요금에 영향을 미치지 않는다는 가설이 기각되었다 (표 6). 각 독립변수들과 종속변수의 관계유무의 t검정 결과 주행거리(X1) 와 전동차 전력(X2)는 전력요금에 영향을 미쳤다 ($p<0.05$).
그러나 부대시설전력은 전력요금에 영향을 미치지 않았다 ($p>0.05$). 중회귀모형은 Y=107596226+4.36X1+2.26X2-7.72X3로 추정되었다.
결정계수 R2 값은 0.905, 조정된 결정계수값은 0.870 이었다.

3-2. 엑셀 통계분석 결과

엑셀 통계분석의 회귀분석 메뉴를 적용하여 분석한 결과는 표 8에 있다. 분산분석결과는 표 6의 매크로 적용 결과와 일치하였다 ($p=0.000189$). 회귀계수의 t 검정 결과들도 표 7의 매크로 적용결과들과 일치하였다. 결정계수 등의 결과들도 매크로 적용 결과와 엑셀 메뉴 결과들과 모두 일치하였다.
중회귀 모형 추정 결과도 일치하였다.

4. 결론

변수들간의 관계성 유무 검정 및 모형 추정에 적용되는 회귀분석은 자료가 쌍을 이루고 있으므로 모든 통계량 계산은 행렬에 의하여 수행된다.
따라서 역행렬 등 행렬 연산을 손쉽게 수행할 수 있는 엑셀에서 쉽게 수행될 수 있다.
또한 회귀 분석에서는 모형 설정이 중요하므로 엑셀의 편리한 기능이 매우 도움이된다. 전문 통계 분석 패키지의 회귀분석 프로그램은 정해진 결과들만이 출력되고 전문적인 통계량들이 많이 포함되어 단순한 자료분석에는 불필요한부분들이 많다. 반면에 엑셀의 함수들로 매크로를 개발하여 회귀 분석을 시행하면 그래픽 메뉴를 활용하여 변수 관계를 가장 잘 설명하는 모형을 찾을 수 있고, 행렬 연산 함수로부터 필요한 결과들만을얻을 수 있어 결과 해설이 용이하다.
비선형 회귀분석도 변수의 관계를 선형이되도록 변형시켜 본 연구에서 개발된 매크로를 적용하여 쉽게 수행될 수 있다.

참고문헌

[1]htttp://www.refee.com/search/?rq=11&qt=%C5%EB%B0%E8%C7%D0%B0%B3%B7%D0 통계학 개론 2008.11.18

[2]http://blog.naver.com/jooning?Redirect=Log&logNo=40056022178엑셀 2007 매크로, 2008.10.16

[3]http://kin.naver.com/detail/detail.php?d1id=1&dir_id=1050202&eid=8IgHCWZFsnzklFN/x/amttnMSz0cPLVV&qb=v6K8v7WlwMzFzbrQvK6x4rTJ&pid=fR822woi5TCssb/pxtosss--480659&sid=SU82WCctT0kAACakqxM 엑셀2007의 통계분석기능, 2007.11.15

[4]http://blog.naver.com/starcandy01?Redirect=Log&logNo=140024543475 중회귀분석 알고리즘

[5] 똑똑한 직원이 숨겨 놓고 혼자 보는 엑셀 함수 전략 <완전개정판> 배남환저 정보문화사 2009