Journal of the Korean Data & Information Science Society
2013, **24**(3), 593–601

http://dx.doi.org/10.7465/jkdi.2013.24.3.593
한국데이터정보과학회지

엑셀 VBA을 이용한 가변수 회귀모형 교육도구 개발

최현석[1] · 박철용[2]

[12]계명대학교 통계학과

접수 2013년 4월 23일, 수정 2013년 5월 16일, 게재확정 2013년 5월 20일

요 약

회귀모형에서 범주형 변수를 독립변수로 포함시켜야 할 경우가 발생한다. 회귀모형의 범주형 변수는 가변수를 통해 수량화된다. 이 연구에서는 하나의 양적 독립변수와 하나 혹은 두 개의 범주형 독립변수를 가지는 회귀모형에 대해 가설검정 결과와 함께 회귀직선을 보여주는 교육용 도구를 엑셀 VBA (Visual Basic for application)를 통해서 구현한다. 가설검정 결과와 회귀직선은 교호작용이 포함된 모형, 교호작용이 없는 모형 및 가변수가 없는 모형에 대해 단계별로 제공된다. 이 교육도구를 통해 가변수와 교호작용의 의미를 더 쉽게 이해할 수 있으며, 나아가 어떤 모형이 주어진 자료에 가장 적합한지 그림을 통해 판단할 수 있게 된다.

주요용어: 가변수, 교호작용, 엑셀 매크로

1. 서론

회귀분석에서 독립변수들은 원칙적으로 구간척도 혹은 비율척도와 같은 양적변수이어야 하나, 성별 (남, 여), 흡연여부 (흡연, 비흡연), 치료방법 (A, B, C), 사회계층 (상류층, 중류층, 하류층) 등과 같은 명목척도나 서열척도 변수를 독립변수로 사용해야 하는 경우가 있다. 예를 들어 종속변수인 혈압과 독립변수인 체중과 연령간의 관계를 연구하고자 할 때 독립변수에 질적변수인 성별과 흡연여부를 포함시키는 경우이다.

회귀변수에서 질적변수를 독립변수에 포함시키기 위해서는 가변수 (dummy variable)를 생성시켜야 한다. 가변수는 0과 1의 값을 취하는 변수를 말하는 것으로, 일반적으로 n개의 범주 (category)가 있는 질적변수에 대해서는 $n-1$개의 가변수를 생성하여 질적변수를 수량화한다. 다시 말해, 독립변수에 명목척도나 서열척도 같은 질적변수를 사용하기 위해서는 0 또는 1로 수량화하여 질적변수의 범주 간 차이를 식별하기 위해 가변수를 사용하는 것이다.

가변수를 이용하여 회귀모형을 사용할 때는 0 또는 1로 수량화 하는 것, 전체회귀식을 구한 후 가변수의 집단별로 회귀식을 구하여 그림을 표현하는 것 그리고 독립변수들 간의 교호작용 등 여러 단계를 거치기 때문에 번거롭고 복잡하다.

많은 통계패키지에서 가변수를 포함한 분석을 별도로 제공되는 것이 아니고 연구자가 질적변수의 범주를 수량화하여 분석하여야 하는 경우가 많다. 범주의 수량화와 교호작용 (interaction) 등을 무시하고 분석을 진행하는 경우, 집단별 회귀직선 (regression line)의 표현 등에 어려움이 발생한다. 따라서 질적변수의 수량화를 진행해주고, 다른 설명변수와의 교호작용을 고려한 모형을 설정하여 검정결과와 회귀

[1] (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 초빙조교수.
[2] 교신저자: (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 교수.
E-mail: cypark1@kmu.ac.kr

직선을 한꺼번에 보여주는 교육도구가 개발된다면 편리할 것이다. 추가적으로 교호작용이 제거되었을 때의 검정결과와 회귀직선, 가변수가 제거되었을 때의 검정결과와 회귀직선을 연동하여 보여줄 수 있다면 가변수와 교호작용의 이해에 많은 도움이 될 수 있을 것이다.

우리가 쉽게 접할 수 있는 엑셀의 VBA (Visual Basic for application) 기능을 이용한 도구들이 계속 개발되고 있다. 서울대학교 통계학과에서는 엑셀 ADD-IN 프로그램인 KESS (Korean educational statistics software)을 개발하였으며 Kim (2012)은 엑셀 매크로기능을 이용한 DES (data encryption standard)의 라운드 키 생성 방법을 연구하였다. 또한 Choi와 Kim (2010)은 엑셀의 함수를 이용한 표본추출에 관하여 연구하였으며, Choi와 Ha (2012)는 엑셀 매크로기능을 이용하여 베이즈 (Bayes) 정리 교육도구를 개발하였다. 앞으로도 엑셀의 VBA 기반 다양한 교육도구들이 계속 개발되어 일반인들이 통계학의 기본개념과 분석 절차를 이해하는 데 많은 도움을 제공할 수 있으리라 기대된다.

본 연구에서는 엑셀의 VBA를 이용하여 하나 혹은 두 개의 질적변수가 있는 경우에 가변수를 이용한 회귀모형의 검정결과와 회귀직선을 단계적으로 보여주는 매크로를 소개한다. 구체적으로 교호작용이 포함된 모형, 교호작용이 제거된 모형 및 가변수가 제거된 모형에서 단계적인 검정결과와 회귀직선을 보여줄 수 있다. 2절에서는 엑셀의 VBA에 기초한 간략한 연구방법을 소개하고 있으며, 3절에서는 프로그램의 구성 및 설명이 나와 있다. 마지막으로 4절에서는 본 연구의 결과를 정리하고 있다.

2. 연구방법

가변수에 대한 설명과 질적변수 범주의 수량화 과정, 교호작용, 모형적합통계량, 각 회귀계수의 통계량, 산점도, 집단별 회귀식과 그래프, 결과분석 등을 단계적 학습이 될 수 있도록 나타내기 위하여 엑셀 VBA를 이용하여 프로그램을 작성하였다. 구체적으로 이 프로그램에서는 프로그램 제어와 함수 사용, 설명, 그래프 등을 위하여 양식도구, 매크로, VBA를 사용하였다. 양식도구는 Dialog Sheet 상에서 대화상자를 사용자가 직접 작성할 때 사용하는 것이고, 매크로는 이용자의 작업 처리를 코드로 기록해 두었다가 나중에 이 코드로 작업을 자동으로 수행하기 위한 명령이다. 엑셀매크로의 경우 반복적으로 실행되는 명령어를 모아서 한 번에 실행하는 것 외에 비주얼 베이직 (Visual Basic)이라는 프로그램언어를 내장하여 단순한 명령어 나열이 아니라 명령어를 이용한 프로그래밍이 가능하도록 되어 있다 (Jacobson, 2002; Walkenbach, 2004).

엑셀의 분석도구는 제한적이어서 함수를 이용한 일반 통계분석에는 한계가 있으나 엑셀 VBA로 프로그램을 작성하면 고급기법들을 사용할 수 있다. 이미 엑셀의 비주얼베이직 매크로를 이용한 다양한 통계교육 혹은 분석도구가 개발되어 활용되고 있다 (Lee, 2008; Choi와 Ha, 2011). 또한 웹사이트 www.unistat.com, www.xlstat.com에서도 엑셀을 이용한 다양한 통계 분석방법을 제공하고 있다.

본 교육용 도구 개발을 위해 VBA Project의 모듈 창에 다음과 같이 코드를 작성하였다.

첫째, Sub 문으로 지정된 셀에서 입력받은 값을 수식을 이용하여 기록하는 프로시저를 작성하였다.

둘째, 엑셀 자체에서 제공되는 분석기능과 VBA 등으로 작성된 프로그램을 연결하였다.

셋째, 명령단추 (command button)를 사용하여 단추를 누르면 바로 매크로가 실행되게 하였다 (Choi와 Kim, 2010).

3. 프로그램의 구성 및 설명

본 연구에서는 양적 종속변수가 하나 있고 질적 독립변수가 한 개나 두 개인 경우의 회귀모형에 적용될 수 있다. 구체적으로 지금까지 개발된 형태는 한 화면에 보여줄 수 있는 결과물의 제한 때문에 한 개의 질적 독립변수의 범주가 두 개나 세 개인 경우, 두 개의 질적 독립변수의 범주가 각각 두 개나 세 개

인 경우로 한정하였다. 구체적으로 이 프로그램에서 사용할 수 있는 가변수 회귀모형의 예제를 살펴보면 Table 3.1과 같다.

Table 3.1 Examples of regression models with dummy variables considered in our program

Independent variables	Regression models with dummy variables
weight (quantitative) sex (male, female)	1) Full model: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon_i$ x_1 : weight $x_2 = \begin{cases} 1, & \text{male} \\ 0, & \text{female} \end{cases}$ 2) Model with dummy: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$ 3) Model without dummy: $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$
age (quantitative) treatment method (A, B, C)	1) Full model: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon_i$ x_1 : age $x_2 = \begin{cases} 1, & \text{A} \\ 0, & \text{B or C} \end{cases}$ $x_3 = \begin{cases} 1, & \text{B} \\ 0, & \text{A or C} \end{cases}$ 2) Model with dummy: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$ 3) Model without dummy: $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$
age (quantitative) sex (male, female) treatment method (A, B, C)	1) Full model: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 + \beta_9 x_2 x_4 + \epsilon_i$ x_1 : age $x_2 = \begin{cases} 1, & \text{male} \\ 0, & \text{female} \end{cases}$ $x_3 = \begin{cases} 1, & \text{A} \\ 0, & \text{B or C} \end{cases}$ $x_4 = \begin{cases} 1, & \text{B} \\ 0, & \text{A or C} \end{cases}$ 2) Models with interaction: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \epsilon_i$ $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \epsilon_i$ 3) Models with dummy(s): $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon_i$ $y_i = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \epsilon_i$ $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$ 4) Model without dummy: $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$

첫 번째 예제를 보다 자세히 설명하도록 하자. 첫 번째 예제는 혈압 y를 종속변수로 하고 체중 x_1와 성별의 가변수 x_2를 독립변수로 포함하는 모형으로 교호작용을 포함한 회귀모형을 다음과 같이 나타낼 수 있다.

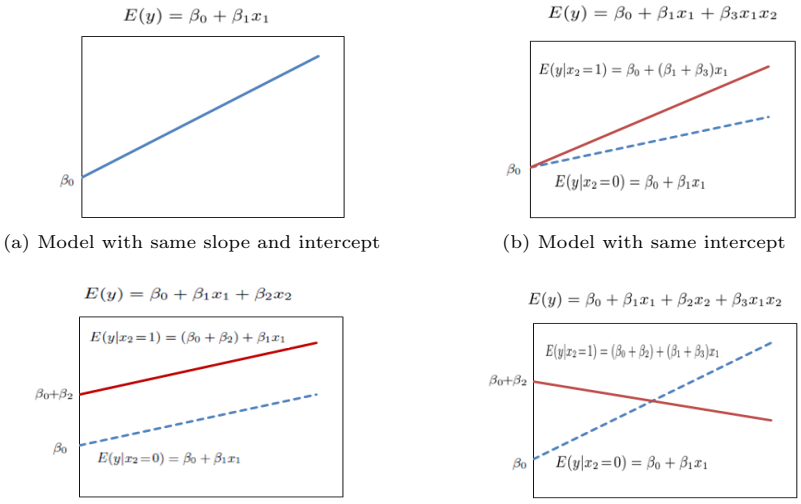
$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon_i \tag{3.1}$$

$$x_2 = \begin{cases} 1, \text{남} \\ 0, \text{여} \end{cases}$$

식 (3.1)은 교호작용까지 포함된 완전모형 (full model)으로 Figure 3.1(d)에서 보는 것과 같이 성별에 따라 서로 다른 회귀직선을 가지게 된다. 자료를 식 (3.1)에 적합시킨 후 각 회귀계수에 대한 검정을 실시한다. $\beta_3 = 0$인 경우는 $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$로 Figure 3.1(c)에서 보는 것과 같이 두 집단

의 회귀직선의 절편은 다르고 기울기가 같게 된다. $\beta_2 = 0$인 경우는 $y_i = \beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2 + \epsilon_i$로 Figure 3.1(b)에서 보는 것과 같이 두 집단의 회귀직선의 절편이 같게 된다 (이 모형은 교호작용은 있는데 주효과 x_2가 없어 위계모형 (hierarchical model)에서는 흔히 고려되지 않지만 회귀직선의 절편이 같게 되는 의미가 있어 포함시켰다). $\beta_2 = \beta_3 = 0$인 경우는 $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$로 Figure 3.1(a)에서 보는 것과 같이 두 집단이 동일한 회귀직선을 가진다.

이 연구에서 제공하려고 하는 교육프로그램은 Figure 3.1에 주어진 전 과정의 회귀직선과 더불어 가설검정의 결과를 단계적으로 보여주는 프로그램이다.



(a) Model with same slope and intercept

(b) Model with same intercept

(c) Model with same slope

(d) Full model

Figure 3.1 Regression lines for full model and for models with or without dummy variables

엑셀을 실행하면 ADD IN 프로그램으로 구성되어 있어 메뉴표시줄에 '가변수 회귀모형 교육'이 추가된다 (Figure 3.2). 각 메뉴를 클릭하면 VBA Project의 모듈 창에서 비주얼 베이직 (Visual Basic)으로 작성한 프로그램의 해당 메뉴로 이동한다. 양식도구를 사용하여 만들어진 메뉴화면 (Figure 3.3)의 각 명령단추를 클릭하여도 VBA Project의 모듈 창에서 비주얼 베이직으로 작성한 프로그램의 해당 메뉴로 이동할 수 있다

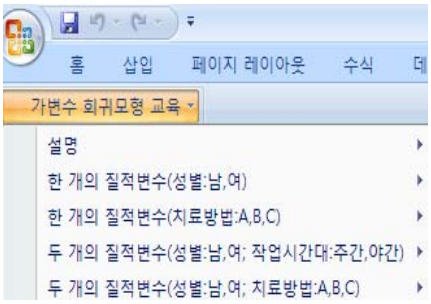


Figure 3.2 Initial display

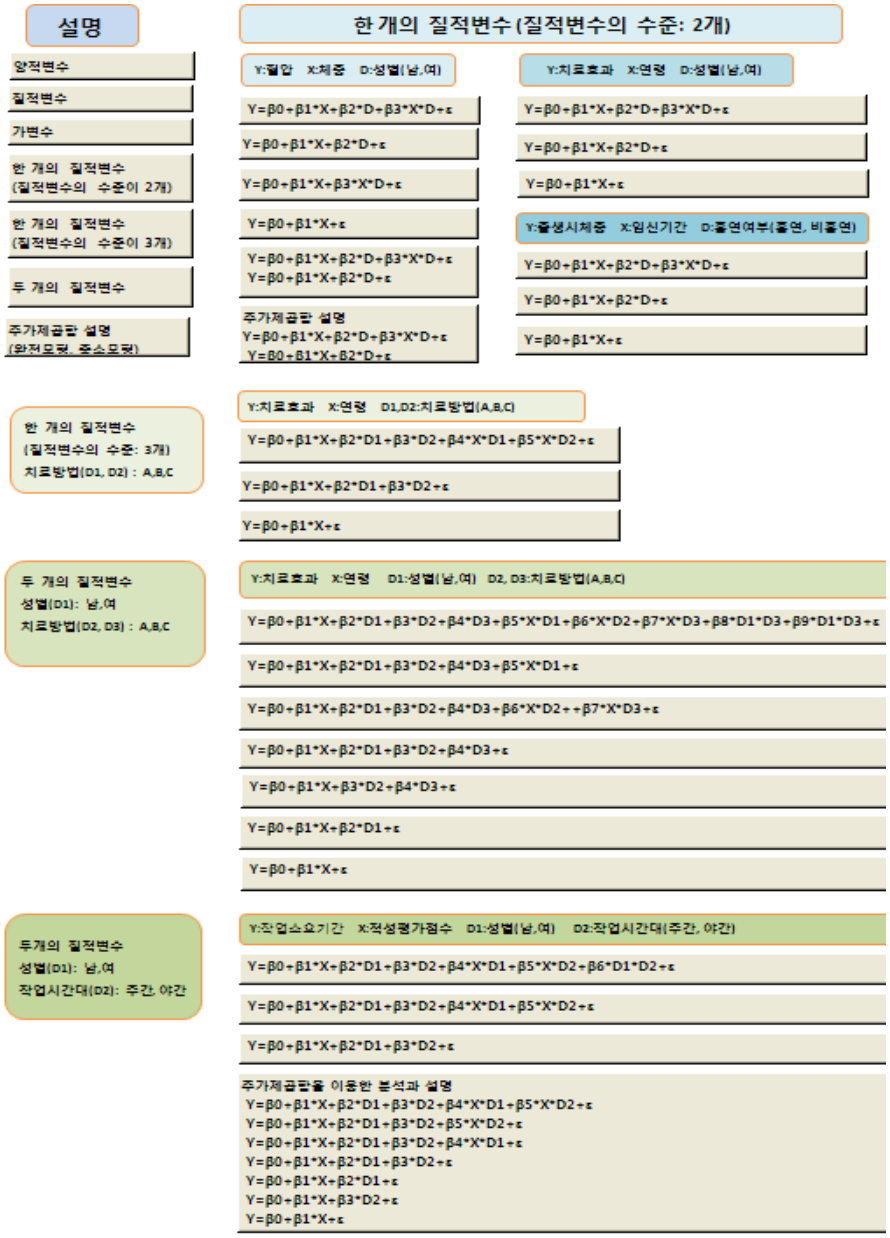


Figure 3.3 Main menu

특정 메뉴가 작동하는 원리를 한 가지 소개하면 다음과 같다. Figure 3.3의 주 메뉴 (main menu)에서 Figure 3.4의 단추를 누르면 Figure 3.5(a)가 나타난다. 이 단추를 누르면 단계적으로 자료와 자료에 대한 설명이 Figure 3.5(b)와 같이 나타나고, 범주를 0과 1로 수량화하는 과정과 교호작용에 대한 자료가 Figure 3.5(c)와 같이 만들어진다. 각 회귀계수에 대한 통계량, 완전 회귀식, 집단별 회귀식과 설명 등이 Figure 3.5(d)와 같이 차례대로 나타난다.

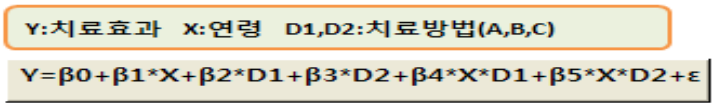
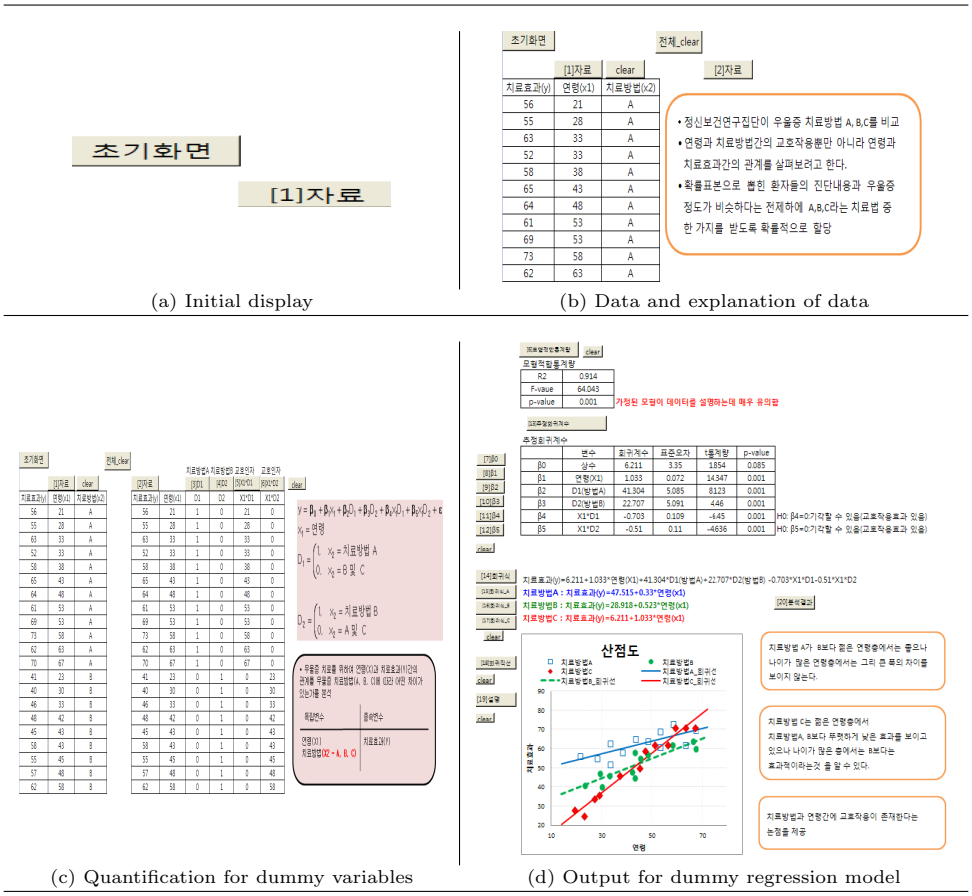


Figure 3.4 Button for an independent variable with three categories



(a) Initial display (b) Data and explanation of data

(c) Quantification for dummy variables (d) Output for dummy regression model

Figure 3.5 Output for the independent variable with three categories

Figure 3.3의 주 메뉴에서 Figure 3.6의 단추를 누르면 Figure 3.7과 같이 교호작용이 있는 완전모형과 교호작용이 없고 가변수만 있는 모형의 분석결과와 회귀직선의 그래프를 제공한다.

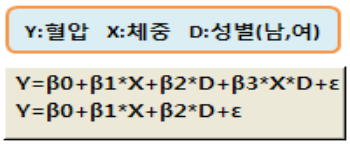


Figure 3.6 Button for an independent variable with two categories

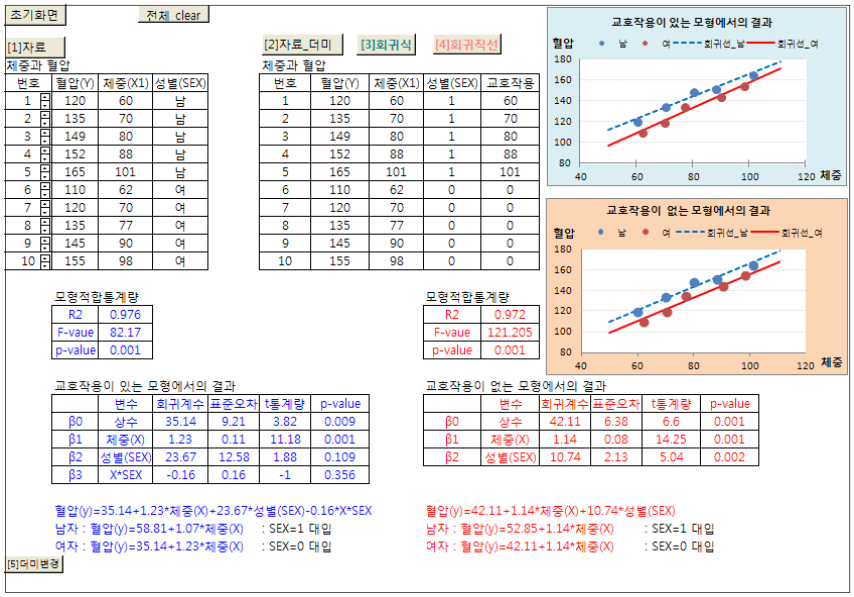


Figure 3.7 Output for the full model and the model without interaction: Two categories

추가적으로 스핀단추를 이용하여 자료값을 변화시킬 수 있는 기능이 있다. 예를 들어 Figure 3.8에서와 같이 혈압의 값을 변화시키면서 교호작용이 없이 가변수만 있는 모형과 교호작용이 있는 모형의 회귀직선과 회귀계수의 통계량 변화를 관찰할 수 있다. 이것을 통해 교호작용이 있는 모형을 교호작용이 없는 모형으로 분석할 때 회귀계수가 많이 달라져 분석에 오류를 범하게 된다는 것을 스스로 체득할 수 있게 해준다.

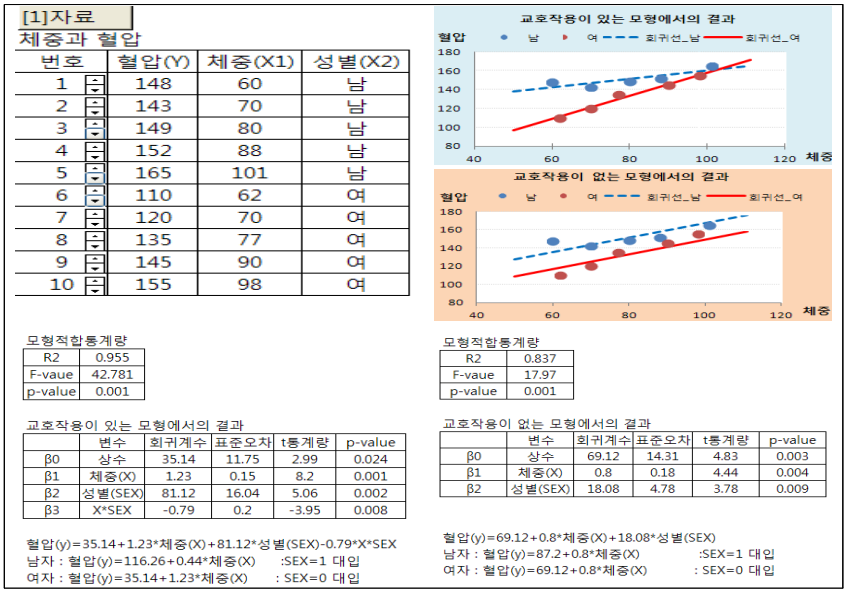


Figure 3.8 Output for the data values changed by spin button: Two categories

4. 결론

회귀모형에서 범주형 변수를 독립변수로 포함시켜야 할 경우 가변수를 통해 수량화한다. 이 연구에서는 하나의 양적 독립변수와 하나 혹은 두 개의 범주형 독립변수를 가지는 회귀모형에 대해 가설검정 결과와 함께 회귀직선을 보여주는 교육용 도구를 엑셀 VBA를 통해서 구현하였다. 가설검정 결과와 회귀직선은 교호작용이 포함된 모형, 교호작용이 없이 가변수만 있는 모형 및 가변수도 없는 모형에 대해 단계별로 제공되고 있다. 이 교육도구를 통해 가변수와 교호작용의 의미를 더 쉽게 이해할 수 있으며, 나아가 어떤 모형이 주어진 자료에 가장 적합한지 그림을 통해 쉽게 판단할 수 있게 도와준다.

이 프로그램의 구동과정은 간단하다. 엑셀을 실행하면 ADD IN 프로그램으로 구성되어 있어 메뉴표시줄에 '가변수 회귀모형 교육'이 추가 된다. 각 메뉴를 클릭하면 VBA Project의 모듈 창에서 비주얼 베이직 (Visual Basic)으로 작성한 프로그램의 해당 메뉴로 이동한다. 양식도구를 사용하여 만들어진 메뉴화면의 각 명령단추를 클릭하여도 VBA Project의 모듈 창에서 비주얼 베이직으로 작성한 프로그램의 해당 메뉴로 이동할 수 있다.

이 프로그램은 통계학전공자나 비전공자 모두 학습자 스스로 통계학습을 할 수 있을 뿐만 아니라 통계학개론 강의에서 보조 자료로도 사용가능하다. 결과가 나오기까지 과정이 설명과 함께 주어지므로 학습효과를 최대화 할 수 있으며, 엑셀 프로그램만 있으면 바로 실행하여 원리와 과정을 학습할 수 있는 장점이 있다. 한 화면상에서 계산과정과 결과를 나타내도록 하기 위하여 범주수가 2개나 3개인 것에 대하여 프로그램을 개발하였으나, 기술적으로 범주수가 많아지더라도 프로그램을 확장하는데 큰 문제가 없다.

References

Choi, H. S. and Ha, J. (2011). Development of process-oriented education tool for Statistics with Excel Macro. *Journal of the Korean Data & Information Science Society*, **22**, 643-650.

Choi, H. S. and Ha, J. (2012). Development of Bayes' rule education tool with Excel Macro. *Journal of the Korean Data & Information Science Society*, **23**, 905-912.

Choi, H. S. and Kim, T. Y. (2010). A study on sampling using the function of excel. *Journal of the Korean Data & Information Science Society*, **21**, 481-491.

Jacobson, R. (2002). *Microsoft excel 2002 visual basic for applications step by step*, Microsoft Press, Redmond, WA.

Kim D. (2012). On the development of DES round key generator based on Excel Macro. *Journal of the Korean Data & Information Science Society*, **23**, 1203-1212.

Lee, J. Y. (2008). An example of participatory statistics class using Excel Macro. *The Korean Journal of Applied Statistics*, **21**, 355-359.

Walkenbach, J. (2004). *Excel 2003 power programming with VBA*, Wiley Publishing, New York.

Journal of the Korean Data & Information Science Society
2013, **24**(3), 593–601

http://dx.doi.org/10.7465/jkdi.2013.24.3.593
한국데이터정보과학회지

An educational tool for regression models with dummy variables using Excel VBA

Hyun Seok Choi[1] · Cheolyong Park[2]

[12]Department of Statistics, Keimyung University

Received 23 April 2013, revised 16 May 2013, accepted 20 May 2013

Abstract

We often need to include categorial variables as explanatory variables in regression models. The categorial variables in regression models can be quantified through dummy variables. In this study, we provide an education tool using Excel VBA for displaying regression lines along with test results for regression models with a continuous explanatory variable and one or two categorical explanatory variables. The regression lines with test results are provided step by step for the model(s) with interaction(s), the model(s) without interaction(s) but with dummy variables, and the model without dummy variable(s). With this tool, we can easily understand the meaning of dummy variables and interaction effect through graphics and further decide which model is more suited to the data on hand.

Keywords: Dummy variable, Excel VBA, interaction effect.

[1] Assistant professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea. E-mail: cypark1@kmu.ac.kr