제13장 회귀분석을 통한 인과관계검증

| 도입 사례 |

미국의 저명한 신문사인 New York Times (NYT)지에서는 대표성 있는 2,500가구를 대상으로 '일간신문구독(interest in reading daily newspaper)'을 결정하는 요인이 무엇인지에 대한 연구를 진행한 바 있습니다. 그들의 연구에 따르면 일간신문구독에 관심을 갖게 하는 가장 중요한 요인은 '세상에 대한 통찰력(sensitive insight into the world)'이고, 신문은 지성, 품위 및 신뢰를 반영해야 한다고 함을 밝혀냈습니다. NYT에서는 회귀분석을 통해 통찰력, 지성, 품위 및 신뢰가 일간신문구독에 미치는 영향을 분석해냈고, 분석 결과를 기초로 2,000만불 규모의 이미지 광고를 수행하게 되었습니다. 물론 이미지 광고는 "Expect the World"라는 광고 문안을 중심으로 지성, 품위, 신뢰를 반영하도록 준비되었습니다. 이미지 광고는 성공적이었고 광고를 실행한 첫해에 NYT 발행부수가 대폭적으로 늘어나는 결과를 얻을 수 있었습니다.

생각해 볼 문제

① 우리나라에서 일간신문 구독에 영향을 주는 요인들은 무엇일까요?

② 일간신문 구독과 영향요인들간의 관계는 어떻게 분석할 수 있을까요?

1. 회귀분석이란 무엇인가?

1) 회귀분석(regression analysis)

- 등간척도 이상의 수준으로 측정된 하나 이상의 독립변수와 하나의 종속변수간의 관계를 밝혀주는 대표적인 인과관계분석기법
- 회귀분석기법을 사용한 분석 결과만을 가지고 인과관계(causation)를 '증명'할 수는 없음
- 독립변수가 종속변수를 야기함(=인과관계)을 의미한다기보다는 독립변수의 변화가 종속변수의 변화를 수학적으로 설명한다고 함을 이해할 필요

  ☞ *하나의 종속변수값을 한 개 이상의 독립변수의 값으로부터 설명하고 예측하는 통계기법*

2) **보통최소자승선형회귀**(ordinary least squares linear regression)분석

♦ **최소자승**(least squares)

- 독립변수와 종속변수간의 관계를 설명해주는 식의 오차를 최소화하는 하나의 방법

  ☞ 최소자승법(the method of least squares)은 19세기 수리통계학의 중심 주제

♦ **보통**(ordinary)

☞ 최소자승법도 여러 종류(예: weighted least squares[WLS] 등)로 구분됨

☞ 보통(ordinary)이란 용어는 최소자승법 중 기본모형을 의미한다고 이해할 수 있음

♦ **선형**(linear)

▫ 독립변수와 종속변수간의 관계의 크기가 독립변수의 크기에 비례함을 의미

♦ **회귀**(regression)

▫ Galton의 '평균으로의 회귀(regression to the mean)'라는 표현에서 유래

3) 회귀분석의 기능

♦ 회귀분석은 독립변수와 종속변수간에 관계가 존재하는지를 분석

♦ 회귀분석은 독립변수와 종속변수간의 관계의 크기도 분석

♦ 특정한 독립변수값(들)에 상응하는 종속변수값을 예측

♦ 다중회귀분석의 경우에는 각 독립변수가 종속변수의 변화를 독립적으로 얼마나 설명해 주는지도 분석

☞ 기존의 독립변수 자료를 이용하여 미래의 종속변수의 수준을 예측할 수 있게 됨

※ 최소자승법

◆ 표본의 분포를 모르는 경우(밀도함수가 주어지지 않는 경우)에 적용 가능

- 독립변수와 종속변수를 동등한 정도로 중요하게 취급하지 않는데 이것을 최소자승 추정법의 비대칭성이라고 함
- 그림
- e_i $(= Y_i - \hat{Y})$의 제곱의 합을 최소화
- X와 Y를 동등하게 취급하지 않음, 회귀직선과 X의 관측치간의 편차에는 아랑곳하지 않고 오로지 회귀직선($\hat{Y}$)과 Y의 관측치(Y_i)간의 편차를 최소화하는데 모든 노력을 집중하는 추정방법임
- 최소자승법을 폭넓게 이용하는 이유

  -> 계산이 간편

  -> 근본적인 이유는 최소자승법은 우리의 궁극적 관심인 종속변수를 정밀하게 추정해 주기 때문임*(두 변수 x, y 간의 관계를 분석하는 이유는 두 변수간의 관계를 밝힘으로써, 이 관계식을 이용해 종속변수인 y를 좀더 정밀하게 이해하고자 하기 때문임)*

4) 회귀분석의 유형

(1) 독립변수의 수 기준

♦ 단순회귀분석(simple regression analysis)

▫ 독립변수가 하나인 회귀분석

♦ 다중회귀분석(multiple regression analysis)

▫ 독립변수가 2개 이상인 회귀분석

(2) 독립변수와 종속변수간의 관계의 크기가 독립변수의 크기에 따라 일정하게 변한다고 가정하는지 여부 기준

♦ 선형회귀분석(linear regression analysis)

▫ 독립변수와 종속변수간의 관계의 크기가 독립변수의 크기에 비례한다고 가정하는 회귀분석

♦ 비선형회귀분석(non-linear regression analysis)

▫ 독립변수와 종속변수간의 관계의 크기가 독립변수의 크기에 비례하지 않고 변화한다고 가정하는 회귀분석

예) 二次회귀모형, 三次회귀모형, 指數회귀모형 등

선형회귀분석: 단순회귀분석 (식 13-1)

(a) $Y = \alpha + \beta X + \varepsilon$

단, X는 독립변수, Y는 종속변수,
ε는 확률적 관계를 나타내기 위한 오차항
α와 β는 미지의 상수인 회귀계수

(b) 매출 = 절편 + 회귀계수 × 광고비
= 80억원 + 15.8 × 광고비

비선형회귀분석: 2차 및 3차 회귀모형의 예 (식 13-2)

(a) $Y = \alpha + \beta X^2 + \varepsilon$

(b) $Y = \alpha + \beta X^3 + \varepsilon$

단, X는 독립변수, Y는 종속변수,
ε는 확률적 관계를 나타내기 위한 오차항
α와 β는 미지의 상수인 회귀계수

2. 단순회귀분석의 원리

♦ 회귀분석의 기본적인 목표

▫ 독립변수와 종속변수간의 관계를 가장 잘 나타내주는 식(=**회귀식**: regression equation)을 찾아내는 것

♦ 단순회귀분석의 예([그림 13-1])

▫ [그림 13-1]의 (b)는, 광고비 지출 수준(=X)에 상응하는 매출의 수준(=Y)을 나타내는 X와 Y의 실제 측정값 5쌍(X_A, Y_A; X_B, Y_B; X_C, Y_C; X_D, Y_D; X_E, Y_E)이 A, B, C, D, E로 표현되어 있는 산포도(scatter diagram)임

☞ ㉠, ㉡, ㉢ 혹은 $\hat{Y}_{\neg}$, $\hat{Y}_{\llcorner}$, $\hat{Y}_{\sqsubset}$ 으로 표현된 3개의 가상의 선중 어떠한 선이 독립변수(광고비[=X])와 종속변수(매출[=Y])간의 관계를 가장 잘 나타내 줄까요?

☞ A, B, C, D, E로부터의 거리가 가장 가까운 선을 찾는 것과 동일

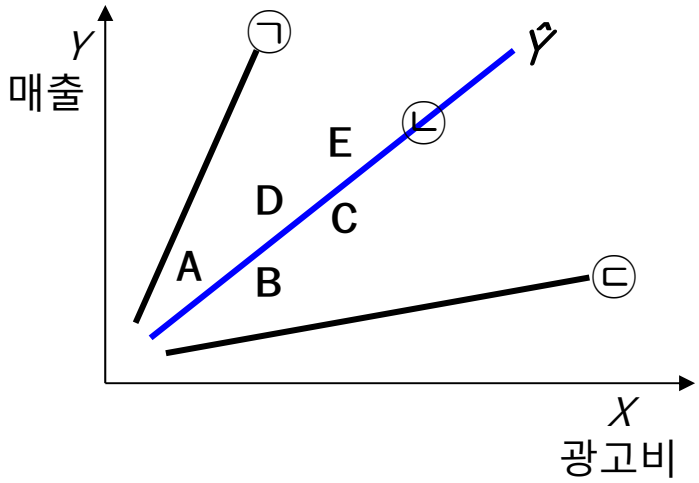
☞ 최적의 가상의 선을 회귀선(regression line)이라고 함

[그림 13-1] 단순회귀분석

(a) $Y = \alpha + \beta X + \varepsilon$

예 : 매출 = 80억 원 + 15.8 × 광고비

(b)



$\min \sum (Y_i - \hat{Y})$ — (가)

$\min \sum |Y_i - \hat{Y}|$ — (나)

$\min \sum (Y_i - \hat{Y})^2$ — (다)

♦ 회귀선을 찾는 3개의 방법

☞ 그림상의 점 하나(= Y_i)로부터 가상의 회귀선(= $\hat{Y}$)까지의 수직(垂直) 거리(vertical distance)인 **잔차**($Y_i - \hat{Y}$)를 공통적으로 사용

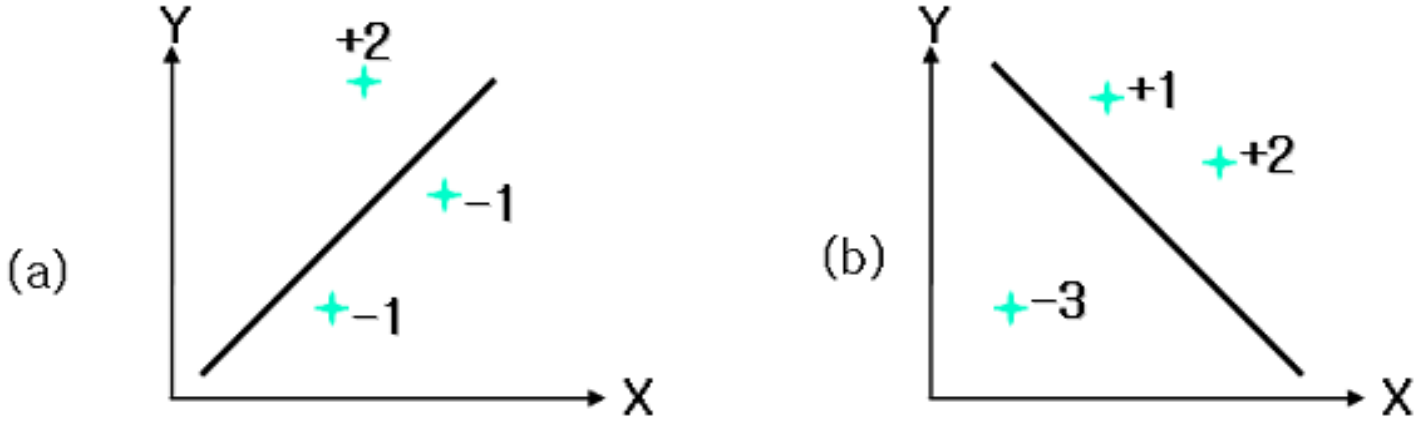
(가) 잔차의 합을 최소화[=min $\sum(Y_i - \hat{Y})$]

▫ 모든 점(즉, A, B, C, D, E)으로부터 가상의 회귀선까지의 거리를[= $\sum(Y_i - \hat{Y})$] 최소화(minimize)하는 방법

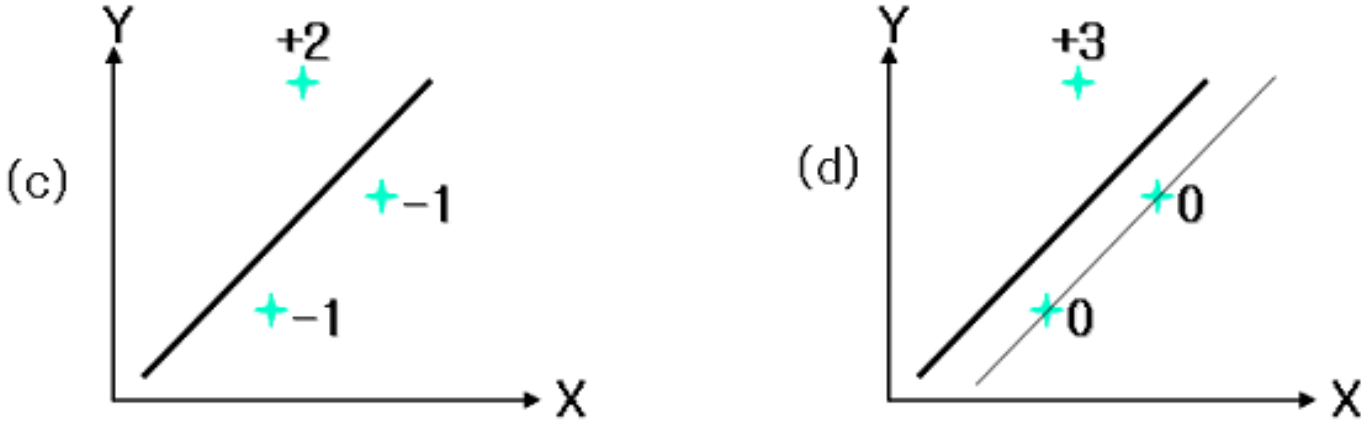
☞ 편차(deviation)의 **단점 보유**: 양의 잔차와 음의 잔차가 우연히 상쇄되어 실제로는 잔차가 있어도 계산상으로만 잔차가 0이 될 수 있음

➪ 동일한 잔차의 합을 가진 가상의 선이 여러 개 생성될 수 있음

[그림 13-2] (A) 잔차의 합이 동일한 회귀선의 예



[그림 13-2] (B) 잔차 절대값의 합과 회귀선



(나) 잔차 절대값의 합을 최소화(= min $\sum|Y_i-\hat{Y}|$)

- (가)의 단점을 극복하고자 개발된 방법
- 잔차의 절대값을 구해 모든 잔차 절대값의 합(= $\sum|Y_i-\hat{Y}|$) 을 최소화(= min $\sum|Y_i-\hat{Y}|$) 하는 방법
  - ☞ 잔차의 절대량이 보존되어 합산됨
  - ☞ 문제점 잔존

  예) [그림 13-2]의 (B)

(다) **각 잔차의 제곱의 합을 최소화**[= min $\sum(Y_i-\hat{Y})^2$]

- (가)와 (나)의 단점을 모두 극복하는 방법
- 각 잔차의 제곱을 구해 그 합[= $\sum(Y_i-\hat{Y})^2$]을 최소화[= min $\sum(Y_i-\hat{Y})^2$]시키는 것
  - ☞ 보통최소자승법(ordinary least squares: OLS)

♦ [그림 13-1](b)의 경우를 예로 들면, 단순회귀분석이란 최소자승법을 이용하여 A, B, C, D, E로부터의 거리가 가장 짧은 회귀선(regression line) 혹은 회귀식(regression equation)을 찾아내는 것을 의미

♦ 회귀계수 (regression coefficient)

- 독립변수가 종속변수를 설명해 주는 크기
  - ☞ 회귀분석이란, 최소자승법을 이용해서 독립변수와 종속변수간의 관계를 설명하는 회귀식의 회귀계수를 찾아내고 회귀식과 회귀계수가 통계적으로 유의한지를 계산해내는 것이 기본 목표임
  - ☞ 결국 최소자승법을 이용한 회귀분석은 min $\sum(Y_i-\hat{Y})^2$ 을 만족시키는 α 와 β 의 특정한 값(즉, 회귀선의 $\hat{\alpha}$와 $\hat{\beta}$)을 구하는 것과 동일한 이야기

단순회귀분석의 회귀계수 계산 (식 13-3)

(a) $\min \sum(Y_i - \hat{Y})^2 = \min \sum[Y_i - (\hat{a} + \hat{b}X_i)]^2$

이 식을 $\hat{b}$ 에 대해 편미분하게 되면

(b) $\hat{b}$ = X, Y 간의 공분산/ X 의 분산 = COV_{XY} / S_X^2

$= [\sum(X_i - \bar{X})(Y_i - \bar{Y})/(N-1)]/[\sum(X_i - \bar{X})^2/(N-1)]$

$= \sum(X_i - \bar{X})(Y_i - \bar{Y})/\sum(X_i - \bar{X})^2$

(c) $\hat{a} = \bar{Y} - \hat{b}\bar{X}$

$\hat{b}$ 은 X와 Y의 공분산을 공평하게 SxSy로 표준화하는 것이 아니라 X의 분산인 S_X^2 만으로 표준화하는 것, 여기서 OLS의 비대칭성의 단면을 또 한번 살펴볼 수 있음

① **비표준화 회귀계수**(unstandardized regression coefficient)

- 원래의 변수들의 측정값을 그대로 사용해서 추정한 회귀계수

② **표준화된 회귀계수**(standardized regression coefficient)

- **베타계수**(beta coefficient 혹은 beta weight)
- 회귀계수의 상대적 중요도를 평가하는 데 유용한 도구
- 독립변수들간의 상대적 중요성을 판단하기 위해 필요한 도구

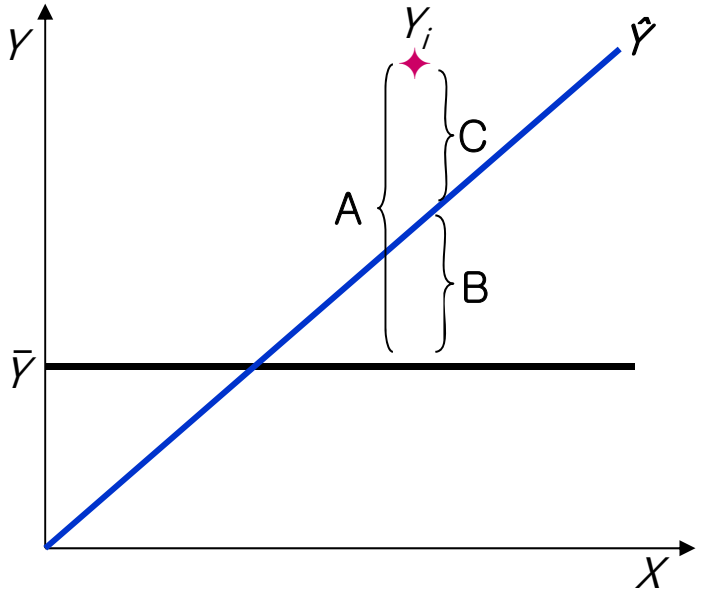
* 표준화

♦ 회귀식의 설명력

▫ **결정계수** (coefficient of determination: R^2) : 0에서 1의 값을 가지고 1에 가까울수록 높은 설명력을 가진 회귀식임

▫ 모든 측정값의 변화폭 중 회귀식에 의해 설명되는 정도

[그림 13-3] 회귀식의 설명력



$$\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2} = \frac{B}{A}$$

결정계수 R^2

결정계수 계산 (식 13-4)

$\Sigma(Y_i - \overline{Y})^2$: Y_i의 총변동폭, **총제곱합**(total sum of squares: ***SST***)

$\Sigma(\hat{Y} - \overline{Y})^2$: **회귀제곱의 합**(regression sum of squares: ***SSR***)

$\Sigma(Y_i - \hat{Y})^2$: **잔차(=오차)제곱의 합**(error sum of squares: ***SSE***)

$SST = SSR + SSE$

$R^2 = SSR / SST$

단, $0 \leq R^2 \leq 1$

3. 다중회귀분석의 원리

♦ 다중회귀분석(multiple regression analysis)

▫ 하나 이상의 독립변수들이 결과변수에 어떠한 영향을 미치는가를 분석해 주는 회귀분석기법

다중회귀분석 모형 (식 13-5)

(a) 기본모형: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$

단, X_1는 첫 번째 독립변수, X_2는 두 번째 독립변수, …, X_n 은 n번째 독립변수
ε는 확률적 관계를 나타내기 위한 오차항
α, β_1, β_2,…, β_n 는 미지의 상수인 회귀계수

(b) 매출 = 60억원 + 15.8 × 광고비 + 5 × 상품품질 + 3 × 판매점 수

(c) 추정된 회귀식: $\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots + \hat{b}_n X_n$

♦ 다중회귀분석의 목표

▫ 각각의 독립변수가 종속변수에 미치는 영향(즉, 회귀계수)을 계산해내는 것

▫ 회귀계수들은 부분회귀계수(partial regression coefficients)

☞ 독립변수들($= X_1, X_2, \cdots, X_n$)이 동시에 한 단위(one unit) 변화하면 종속변수는 ($\beta_1 + \beta_2 + \cdots + \beta_n$)만큼 변화함

♦ 회귀식/계수 추정

▫ 잔차제곱의 합을 최소화[$= \min\Sigma(Y_i - \hat{Y})^2$]하는 조건을 만족시키는 회귀계수의 특정한 값($= \hat{a}, \hat{b}_1, \hat{b}_2, \cdots, \hat{b}_n$)을 추정

☞ 구체적으로는, 편미분을 통해서 모든 잔차제곱을 최소화시키는 방정식들을 구하고, 이 연립방정식(system of simultaneous equations)을 풀어서 회귀계수를 구함

☞ 일반적으로 다른 집단간(across different groups)의 회귀계수 크기비교는 비표준화회귀계수를 사용하고 같은 집단 내에서의 변수들의 상대적 중요성은 표준화회귀계수를 사용

♦ 회귀식의 설명력

① 단순회귀분석시의 회귀식의 설명력(= R^2)

▫ coefficient of determination (결정계수)

② 다중회귀분석시의 회귀식의 설명력(= R^2)

▫ coefficient of multiple determination

③ 다중회귀분석시의 회귀식의 설명력(= adjusted R^2)

▫ **수정된 R^2**

☞ 독립변수의 수와 표본의 크기를 고려한 회귀식의 설명력

수정된 R^2 (식 13-6)

$$\text{수정된 } R^2 = 1-(1-R^2)\frac{N-1}{N-k-1}$$

단, n은 표본의 크기,
k는 독립변수의 수

4. 다중회귀분석시 주의점

♦ **다중공선성(multi-collinearity)의 부재**

▫ **독립변수들간에는 서로 관련이 없어야 한다는 조건**

♦ 다중공선성 판단도구

① **공차한계**(tolerance)

▫ 독립변수들간의 다중상관관계(multiple correlation)를 나타내는 통계량으로 (1 - R_i^2)으로 계산

☞ "공차한계 ≥ 0.1" 면 다중공선성 문제가 없는 것으로 판단

② **분산팽창지수**(variance inflation factor: **VIF**) = 공차한계의 역수

☞ "분산팽창지수 ≤ 10"면 다중공선성 문제가 없는 것으로 판단

♦ **다중공선성 대처방안** (* 가장 중요한 점: 왜 이런 현상 발생하는지 회귀분석모형의 적합성을 관련 이론에 입각해서 분석해야 한다는 점)

① 상관관계가 높은 독립변수들 중 중요하지 않다고 판단되는 변수를 회귀분석에서 제외 (이 선택이 잘못되면 중요 변수의 누락 결과를 가져옴)

② 독립변수들에 대한 요인분석을 실행하여 이들 변수들을 대표하는 요인값 등을 계산하고 이것을 새로운 독립변수로 사용

♦ **선형회귀분석의 기본 가정**

① 종속변수를 독립변수의 선형함수로 나타낼 수 있다고 하는 **선형성(linearity) 가정**

-> 성립 여부는 잔차(오차)의 산포도 분석으로 가능

※ 선형성 가정 위배 시 대처방안

➪ 비선형회귀분석(예: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \varepsilon$) 시도 가능

➪ 혹은 비선형관계(예: $Y = \alpha X^{\beta}$)를 선형관계로 변환(예: $\log Y = \log \alpha + \beta \log X$)후 선형회귀분석 수행가능

② 등분산성(homoscedasticity) 가정

- 모든 잔차의 분산은 일정하다는 가정 (모든 잔차는 평균이 0이고 동일한 분산 가짐) -> 성립 여부는 잔차(오차)의 산포도를 검토함으로써 분석 가능

※ 등분산성가정 위배 시 대처방안 (이분산성; heteroscedasticity)

➪ 보통최소자승법(ordinary least squares) 대신 일반화최소자승법(generalized least squares)이나 가중최소자승법(weighted least squares)을 이용한 계수추정 사용

③ 잔차의 상호독립 가정

- 모든 잔차는 상호독립적이어야 한다는 가정 (자기상관(autocorrelation)이 없어야 함)
  - Durbin-Watson**통계량분석** 추가 : 값이 2에 접근하면 독립성 가정에 큰 문제가 없다고 해석

※ 잔차의 자기상관 존재 시 대처방안

➪ 비선형관계를 다루는 시계열분석 혹은 비선형회귀분석기법 사용

④ 잔차의 정규성(normality) 가정

- 모든 잔차는 평균이 0이고 분산이 σ^2 인 정규분포를 따른다는 가정
  - 판단 방법 : SPSS 정규확률도표를 선택하고 검토-> 잔차항 히스토그램이 좌우대칭인 종의 모양일 것, 표준화 잔차들이 직선 대각선을 형성(<- 잔차들이 정규분포 이루는 경우)

5. 특수회귀분석

1) Dummy변수를 이용한 회귀분석

♦ 독립변수가 명목척도로 측정된 경우에는 명목척도를 구성하는 각 범주의 종속변수에 대한 영향력을 분석할 수 있는 **추가적인 준비**가 필요함

☞ 이와 같은 경우에 사용할 수 있는 특수한 회귀분석방법 중의 하나가 dummy coding(혹은 dummy변수)을 이용한 회귀분석임

♦ Dummy coding

▫ 명목변수를 구성하는 각 범주별로 특정 범주에 속하는 경우에는 1을 부여하고 속하지 않으면 0을 부여하는 coding방법

♦ Dummy변수

▫ Dummy coding 과정을 거쳐 새로 만들어진 변수

♦ Dummy변수의 수

▫ 명목변수를 구성하는 범주(category)의 수–1

☞ $(C-1)$개

Dummy변수의 필요성 (식 13-7)

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3$$

단, $\hat{y}$ 는 매출, X_1는 광고비, X_2는 대리점 수, X_3는 계절

〈표 13−1〉 Dummy변수

명목변수의 범주 \ Dummy변수	X_3	X_4	X_5
봄	0	0	0
여름	1	0	0
가을	0	1	0
겨울	0	0	1

Dummy변수를 이용한 회귀식 (식 13-8)

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \hat{b}_4 X_4 + \hat{b}_5 X_5$$

단, $\hat{Y}$ 는 매출, X_1는 광고비, X_2는 대리점 수, X_3는 여름, X_4는 가을, X_5는 겨울

$\hat{b}_3$는 여름($X_3 = 1$, $X_4 = 0$, $X_5 = 0$)이 매출에 미치는 영향

$\hat{b}_4$는 가을($X_3 = 0$, $X_4 = 1$, $X_5 = 0$)이 매출에 미치는 영향

$\hat{b}_5$는 겨울($X_3 = 0$, $X_4 = 0$, $X_5 = 1$)이 매출에 미치는 영향

* **봄**의 경우는 모든 더미변수의 값이 0이라는 특징을 가지고 다른 계절들과 구분이 됨, 봄은 계절효과를 판단하는 기준(=control) 계절이라 해석 가능

2] 상호작용효과의 회귀분석

♦ **주효과**(main effect)

▫ 독립변수들의 종속변수에 대한 독립적인 효과

♦ **상호작용효과**(interaction effect)

▫ 강화효과와 조절효과를 포함하는 효과

♦ **강화효과**(amplification effect)

▫ 독립변수의 종속변수에 대한 영향력(예: 흡연)이 어떠한 조건(예: 음주와 동시에 흡연)에 의해 더 강해지는 것

♦ **조절효과**(moderation effect)

▫ 독립변수의 종속변수에 대한 영향력(예: 흡연)이 어떠한 조건(예: 음주와 동시에 흡연)에 의해 더 약해지는 것

회귀분석에서의 상호작용효과 (식 13-9)

(a) $\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2$

(b) $\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_1 X_2$

예) <식 13-9> (a)에서 추정된 종속변수(=$\hat{Y}$)를 건강이라고 하고 X_1는 흡연량, X_2는 음주량이라고 가정하면, $\hat{b}_1$는 흡연이 건강에 미치는 독립적인 영향, $\hat{b}_2$는 음주가 건강에 미치는 독립적인 영향

☞ *흡연과 음주를 동시에 하는 경우의 건강에 미치는 영향의 크기는?*

☞ <식 13-9> (b)의 $\hat{b}_3$는 흡연(=X_1)과 음주(=X_2)를 동시에 하는 경우(=X_1X_2)의 영향력

☞ 상호관계를 표현하기 위해 새로 만들어진 독립변수(=X_1X_2)와 기존의 독립변수들(X_1, X_2)간에 존재할 수밖에 없는 다중공선성의 문제가 발생

♦ **상호작용효과분석시 다중공선성 제거방법**

▫ **편차변환**(centering)

▫ 등간척도나 비율척도의 수준으로 측정된 연속(continuous) 독립변수의 경우 상호작용을 표시하는 새 변수(예: $X_1 X_2$)를 구성하기 전에 상호작용을 구성하는 원 변수들(=X_1, X_2)에서 각각의 평균을 차감한 편차점수(= $X_1 - \bar{X}_1$, $X_2 - \bar{X}_2$)를 구성하는 것

♦ **변수의 편차변환 후 회귀분석을 수행**

☞ 변환 후의 회귀계수와 변환전의 회귀계수는 동일

☞ 변환 후의 독립변수들간의 상관관계를 제거

편차변환 후의 상호작용효과 회귀분석 (식 13-10)

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1^* + \hat{b}_2 X_2^* + \hat{b}_3 X_1 X_2^*$$

단, $X_1^* = X_1 - \bar{X}_1$; $X_2^* = X_2 - \bar{X}_2$; $X_1 X_2^* = (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)$

♦ 회귀분석

▫ 표본을 대상으로 관찰한 변수들의 측정값으로 표본통계량을 계산하고, 표본통계량과 해당 표본통계량의 표본분포(sampling distribution)에 근거해서 모집단에서도 예측된 관계가 성립하는지에 대한 가설을 검정

<표 13-2> 다중회귀분석시의 실질가설

연구가설(H_1): 회귀식의 결정계수는 0이 아닐 것이다. (a) 독립변수들은 종속변수를 유의하게 설명할 것이다.
회귀식의 회귀계수 중 하나 이상은 0이 아니다. (b) 적어도 하나 이상의 독립변수는 종속변수에 유의한 영향을 줄 것이다.

귀무가설(H_0): 회귀식의 결정계수는 0일 것이다. (c) 독립변수들은 종속변수를 유의하게 설명하지 못할 것이다.
회귀식의 모든 회귀계수는 동시에 0이다. (d) 어떠한 독립변수도 종속변수에 유의한 영향을 주지 못할 것이다.

<표 13-3> 다중회귀분석의 통계가설

H_0: $R^2_{pop} = 0$(a)

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_n = 0$(b)

♦ <표 13-3>의 (a) 기준 가설검정 내용

▫ H_0: $R^2_{pop} = 0$이라는 가정하에서, 표본에서의 R^2값($= \Sigma(\hat{Y} - \bar{Y})^2 \,/\, \Sigma(Y_i - \bar{Y})^2$)이 R^2_{pop}로부터 얼마나 멀리 떨어져 있는지를 판단하는 것

☞ 표본에서의 R^2값을 이용한 검정통계량(test statistic)을 계산해서 결정계수 관련 가설을 검정하게 됨

다중회귀분석시 결정계수의 검정통계량 (식 13-11)

F = 평균회귀제곱/평균잔차제곱

$$= \frac{SSR/k}{SSE/(n-k-1)}$$

$$= \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

단, SSR 은 회귀제곱의 합(regression sum of squares)

SSE 는 잔차제곱의 합,

n은 표본의 크기,

k는 독립변수의 수,

R^2는 결정계수

♦ SPSS Statistics를 통한 다중회귀분석 실시 결과 검토(<표 13-4>)

♦ 종속변수

▫ '일제상품에 대한 구매의도(변수명 jppi3),'

♦ 독립변수

▫ '전쟁적대감(변수명 aniwar3),' '일제상품품질(변수명 jppq2)' 및 '자문화중심주의(변수명 ethno1)'의 3개

<표 13-4> 회귀분석 결과

진입/제거된 변수(b)

모형	진입된 변수	제거된 변수	방법
1	aniwar3, jppq2, ethno1(a)	.	입력

a 요청된 모든 변수가 입력되었습니다.
b 종속변수: jpi3

모형 요약(b)

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차	Durbin-Watson
1	.383(a)	.147	.135	1.53174	2.024

a 예측값: (상수), aniwar3, jppq2, ethno1
b 종속변수: jpi3

분산분석(b)

모형		제곱합	자유도	평균제곱	F	유의확률
1	선형회귀분석	89.010	3	29.670	12.646	.000(a)
	잔차	518.519	221	2.346		
	합계	607.529	224			

a 예측값: (상수), aniwar3, jppq2, ethno1
b 종속변수: jpi3

계수(a)

모형		비표준화 계수		표준화 계수	t	유의확률	공선성 통계량	
		B	표준오차	베타			공차한계	VIF
1	(상수)	4.157	.850		4.890	.000		
	jppq2	.384	.114	.210	3.361	.001	.993	1.007
	ethno1	−.274	.067	−.255	−4.064	.000	.979	1.021
	aniwar3	−.214	.090	−.148	−2.370	.019	.984	1.016

a 종속변수: jpi3

♦ <표 13-4>의 '진입/제거된 변수'표 검토
▫ 회귀식의 추정에서 실제로 사용된 독립변수 혹은 제외된 변수가 표시되어 있음

♦ 회귀분석 실행시 독립변수들의 실제 투입순서

▫ 회귀분석 실행시(즉, 회귀식의 추정시) 독립변수들의 실제 투입순서가 회귀계수의 추정에 영향을 줄 수 있기 때문에 연구자가 SPSS Statistics의 회귀분석 명령을 수행할 때 독립변수의 회귀식 내 투입방법을 지정할 수 있음

① '동시적 변수투입(enter)'방법

▫ '입력'방법

▫ 모든 독립변수들을 동시에 회귀분석모형에 포함하여 분석

② '단계적 회귀분석(stepwise regression)'방법

▫ 종속변수를 설명하는 데 중요한 독립변수 순으로 회귀식 내에 투입하여 해당 회귀계수를 추정하는 방법

☞ 탐색적인 입장에서나 계수추정의 효율성 입장에서 바람직

③ 제거(remove)법, 전진(forward)법, 후진(backward)법 등

♦ <표 13-4>의 '모형요약'표 검토

▫ 추정된 회귀식의 설명력인 R^2값(=.147)과 수정된 R^2값(=.135)이 제시됨

♦ <표 13-4>의 '분산분석'표 검토

▫ 회귀식의 설명력 R^2이 통계적으로 유의한지 여부를 판단해 주는 결과제시

▫ 즉, H_0: R^2_{pop}=0가 설립하는지 여부를 검정한 결과

☞ 추정된 회귀식의 R^2에 해당하는 F 값(=12.646)의 유의확률은 p =.000

☞ 유의수준 α =0.01에서, 모집단에서의 회귀식의 설명력은 0이라는 귀무가설(H_0: R^2_{pop}=0 혹은 H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_n = 0$)이 안전하게 기각되고 R^2_{pop}이 0보다 유의미하게 큰 수라는 결론을 내릴 수 있음

☞ 즉, <표 13-2>의 (b) "회귀식의 회귀계수 중 하나 이상은 0이 아니다"라는 주장이 성립한다는 결론

☞ 결국 '전쟁적대감,' '일제상품품질' 및 '자문화중심주의'의 3개 독립변수로 구성된 회귀식이 '일제상품에 대한 구매의도'라는 종속변수를 유의미하게 설명해 줌

♦ <표 13-4>의 '계수'표 검토

▫ 종속변수의 변화를 의미 있게 설명해 주는 독립변수는 무엇이고 회귀계수의 크기는 어느 정도인지에 대한 분석

♦ <표 13-4>의 '계수'표에서 비표준화계수 검토
  ▫ 원래 측정된 변수의 측정값(raw score)을 대변하는 독립변수의 종속변수에 대한 영향력
  ▫ 일제상품품질'의 비표준화계수가 가장 큼

♦ <표 13-4>의 '계수'표에서 표준화계수 검토
  ▫ 독립변수들의 상대적인 중요도를 기준으로 한다면 '자문화중심주의'의 영향이 제일 큼

♦ 개별 회귀계수의 통계적 유의성 검토
  ☞ 표본에서 β_i의 검정통계량(test statistic)을 구하여 가설을 검정

<표 13-5> 개별 회귀계수의 통계가설

H_0: $\beta_i \neq 0$
H_0: $\beta_i = 0$

개별 회귀계수의 검정통계량 (식 13-12)

$$t = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

단, β_i는 추정된 회귀계수, $SE_{\hat{\beta}_i}$는 회귀계수 $\hat{\beta}_i$의 표준오차

♦ <표 13-4>의 '계수'표 재검토

▫ 추정된 회귀식의 각 비표준화계수 및 표준화계수에 상응하는 검정통계량 t 값과 이에 해당하는 유의확률

☞ '일제상품품질'이 0.001, '자문화중심주의'가 0.000, 그리고 '전쟁적대감'이 0.019로 제시됨

☞ 유의수준이 α =0.01/α =0.05에서 귀무가설(H_0: β_i =0)이 안전하게 기각되고 회귀계수 β_i는 0과 유의미하게 다른 수라고 하는 결론(단, 전쟁적대감은 유의수준 0.01인 경우는 유의하다고 할 수 없음; 유의수준 0.05인 경우는 유의함)

♦ <표 13-4>의 '계수'표의 '공선성통계량' 검토

▫ 모든 공차한계가 0.1보다 충분히 크고, 모든 분산팽창지수가 10보다 충분히 작음

☞ 독립변수들간에는 다중공선성이 존재하지 않는다는 결론

♦ 회귀분석 대상 자료파일 열기

[그림 13-4] 회귀분석 대상 자료파일 열기의 실행

intmktg.sav [데이터세트1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 확장(X) 창(W) 도움말(H)

표시: 51 / 51 변수

	id	jpnprod1	jpnprod2	jpnprod3	jpnprod4	jpnprod5	jpnprod6	jpnpi1	jpnpi2	jpnpi3	jpnpi4	jpnpi5	jpnpi6	ethno1
1	1.00	5.00	2.00	5.00	5.00	6.00	6.00	1.00	1.00	5.00	2.00	3.00	4.00	2.00
2	2.00	4.00	3.00	4.00	4.00	5.00	5.00	3.00	2.00	3.00	2.00	2.00	2.00	2.00
3	3.00	7.00	3.00	6.00	7.00	7.00	5.00	3.00	3.00	5.00	3.00	7.00	7.00	7.00
4	4.00	5.00	2.00	6.00	7.00	5.00	5.00	2.00	3.00	6.00	2.00	2.00	7.00	3.00
5	5.00	5.00	2.00	7.00	5.00	6.00	7.00	3.00	2.00	2.00	4.00	2.00	3.00	5.00
6	6.00	3.00	2.00	5.00	4.00	4.00	3.00	2.00	1.00	2.00	2.00	1.00	4.00	3.00
7	7.00	6.00	2.00	5.00	7.00	6.00	7.00	1.00	1.00	3.00	2.00	2.00	4.00	1.00
8	8.00	6.00	3.00	5.00	4.00	6.00	6.00	4.00	2.00	2.00	2.00	2.00	4.00	2.00
9	9.00	6.00	2.00	6.00	7.00	6.00	6.00	3.00	2.00	5.00	4.00	2.00	2.00	2.00
10	10.00	6.00	2.00	6.00	6.00	6.00	5.00	2.00	2.00	2.00	4.00	3.00	5.00	3.00
11	11.00	6.00	3.00	5.00	5.00	3.00	4.00	3.00	3.00	4.00	2.00	4.00	2.00	3.00
12	12.00	5.00	2.00	5.00	6.00	5.00	6.00	3.00	5.00	6.00	4.00	4.00	7.00	6.00
13	13.00	4.00	2.00	5.00	6.00	6.00	6.00	4.00	4.00	5.00	4.00	2.00	5.00	5.00
14	14.00	5.00	2.00	7.00	5.00	4.00	5.00	4.00	2.00	4.00	4.00	4.00	6.00	3.00
15	15.00	7.00	2.00	5.00	7.00	5.00	5.00	2.00	2.00	2.00	4.00	2.00	2.00	4.00
16	16.00	5.00	4.00	6.00	7.00	5.00	5.00	4.00	4.00	5.00	2.00	2.00	3.00	4.00
17	17.00	5.00	2.00	5.00	4.00	3.00	4.00	2.00	4.00	4.00	4.00	2.00	6.00	2.00
18	18.00	6.00	3.00	6.00	6.00	5.00	5.00	2.00	1.00	3.00	3.00	2.00	4.00	2.00
19	19.00	7.00	3.00	6.00	6.00	6.00	.	4.00	3.00	3.00	2.00	4.00	2.00	3.00
20	20.00	7.00	1.00	7.00	7.00	6.00	7.00	5.00	2.00	3.00	4.00	3.00	5.00	5.00

데이터 보기 변수 보기

IBM SPSS Statistics 프로세서 준비 완료 Unicode:ON

♦ 필요시 **역코딩** 실행

▫ 설문문항의 질문방향을 모두 동일하게 구성하지 않는 경우 필요

☞ 응답자들이 각 설문문항을 잘 읽고 응답하는지를 체크하기 위한 수단으로 일부 설문문항을 일부러 부정적(혹은 긍정적)으로 구성

예) 3개의 문항으로 일제상품의 구매의도를 측정하는 경우에 2개의 문항은 "나는 기회가 되면 일제상품을 구매할 것이다"라는 형식의 긍정적인 질문으로 구성하고, 3번째 문항은 "가능한 한 나는 일제상품의 구매를 피할 것이다"라는 부정적인 형식의 질문을 구성

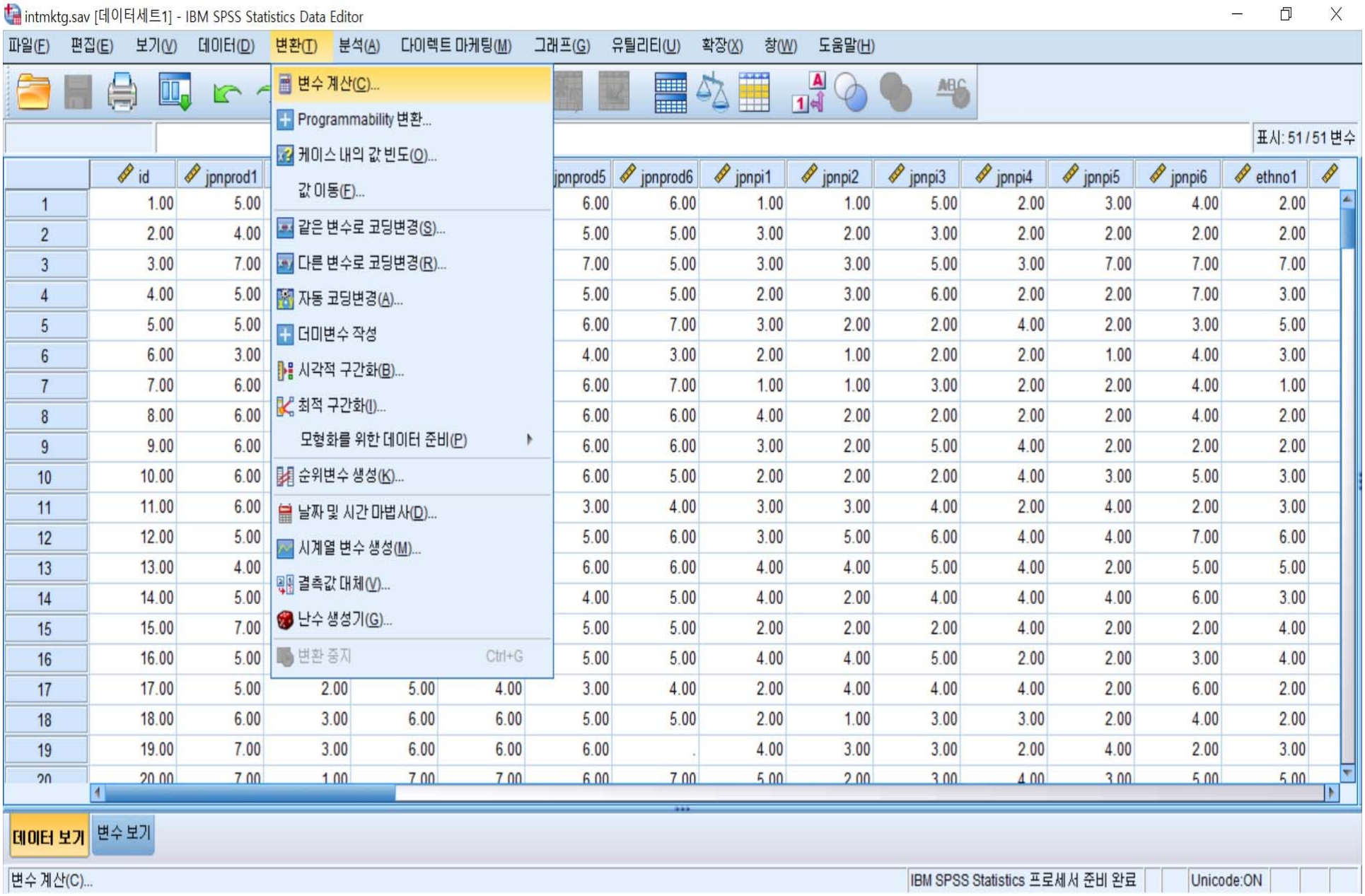
♦ 역코딩(reverse coding)

▫ 동일 현상에 대해 측정문항들에 대한 응답의 예상 방향이 다를 경우에, 각종 통계분석을 수행하기 전에 동일한 방향으로 수정해주는 작업

♦ [그림 13-5]에서는 역코딩을 통한 새 변수생성

☞ [그림 13-5]에서와 같이 **변환(C)** → **변수계산(C)**을 차례로 선택

[그림 13-5] 역코딩을 통한 변수계산



♦ [그림 13-5]에서와 같이 **변환(C)** → **변수계산(C)**을 차례로 선택하게 되면, [그림 13-6]에서와 같은 '변수계산'창이 뜨게 됨

☞ 이 창에서 역코딩으로 생성될 새로운 변수의 이름을 '대상변수(T)' 창에 입력

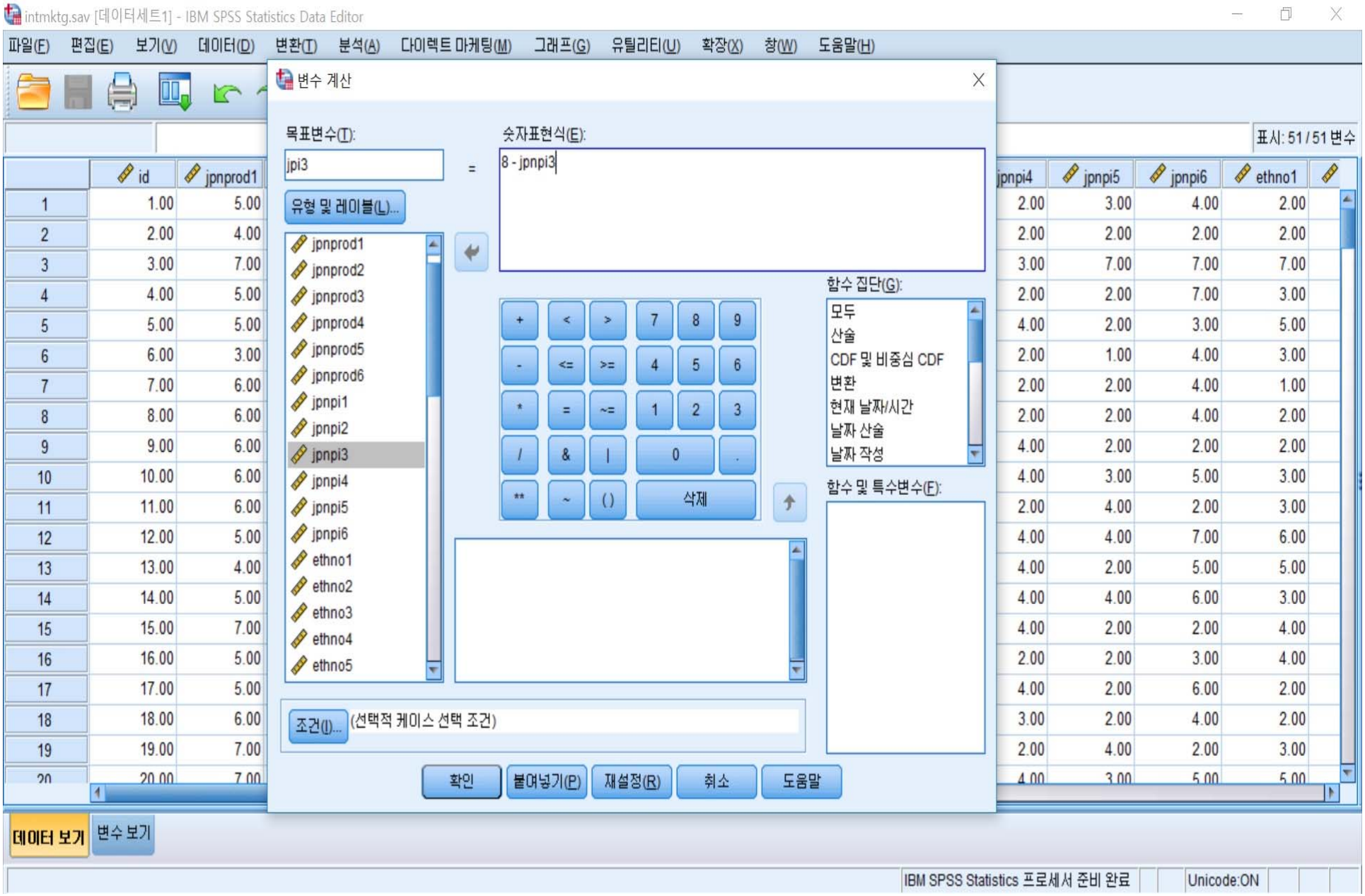
예) 아래에서는 jpi3라고 입력되어 있습니다.

☞ 오른쪽의 '숫자표현식(E)' 창에는 역코딩 대상변수와 역코딩의 규칙을 입력

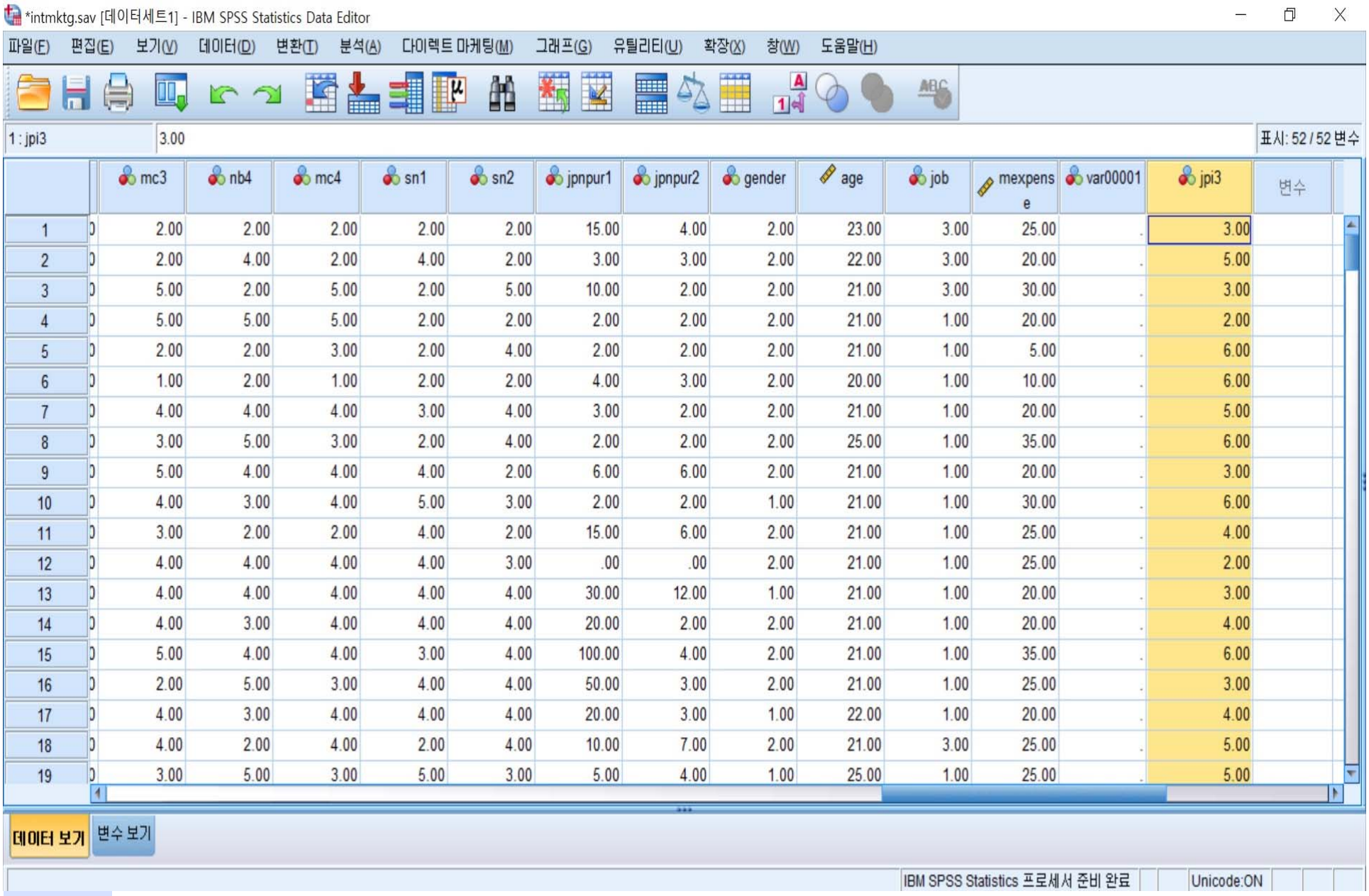
▫ 역코딩에서 사용하는 기본적인 규칙은 "새 변수=[(역코딩 대상 변수의 최대 응답 범주+1)–역코딩 대상 변수]"

♦ [그림 13-6]에서와 같은 준비가 끝나면 **확인** 탭을 눌러 변수계산을 실행

[그림 13-6] 역코딩을 통한 새 변수의 정의 1

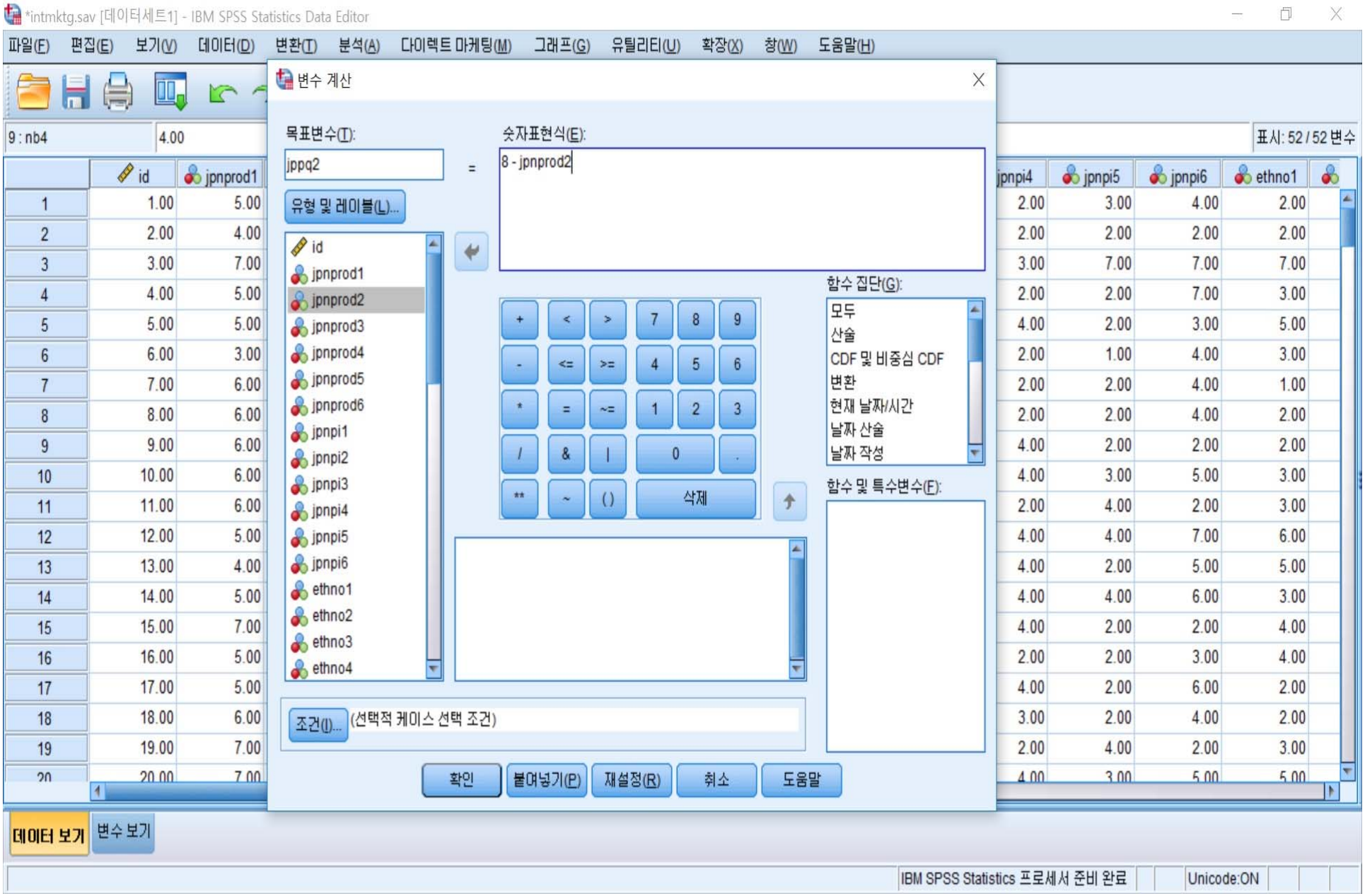


[그림 13-7] 역코딩을 통한 새 변수의 생성 1

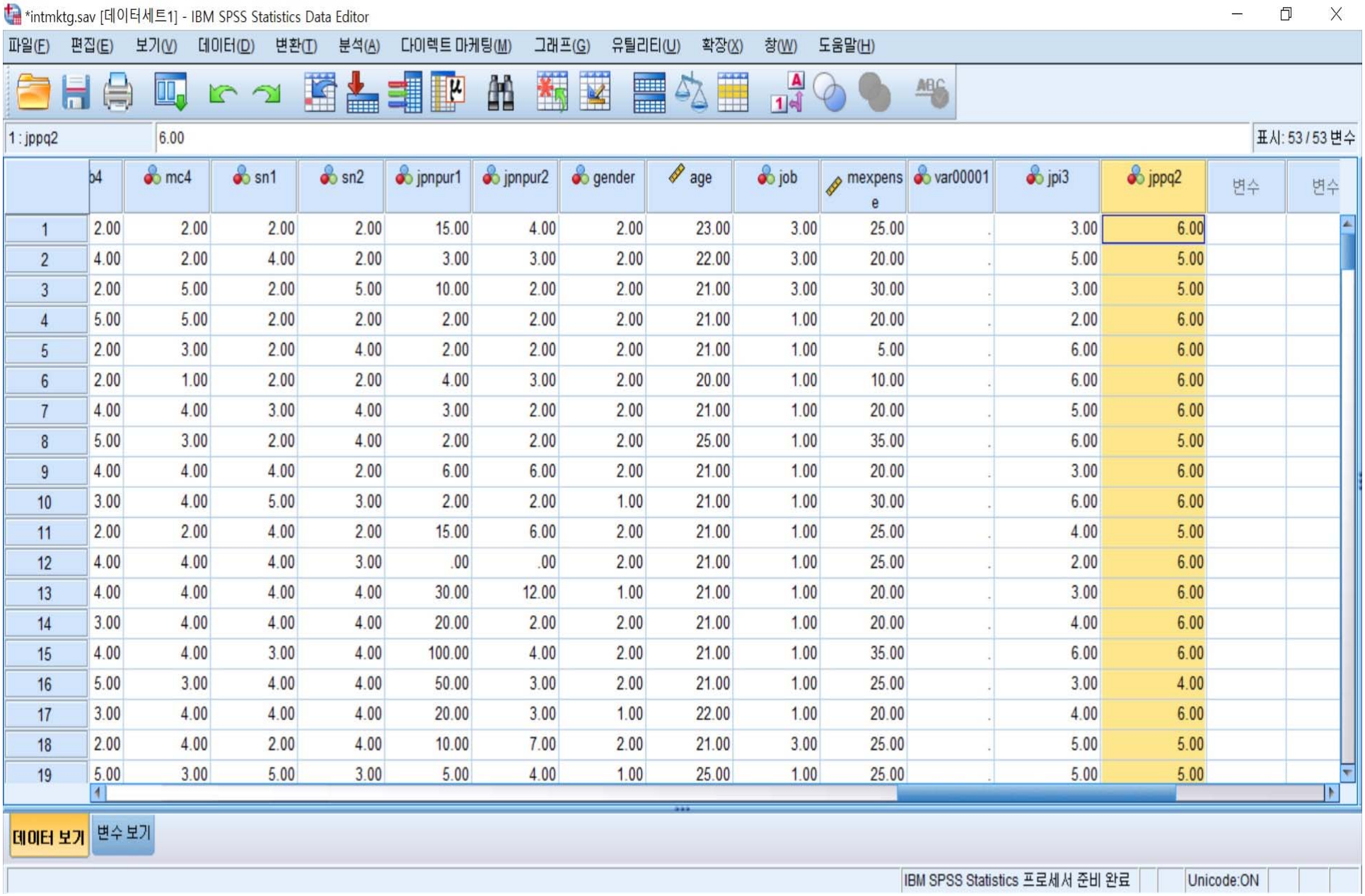


	mc3	nb4	mc4	sn1	sn2	jpnpur1	jpnpur2	gender	age	job	mexpense	var00001	jpi3	변수
1	2.00	2.00	2.00	2.00	2.00	15.00	4.00	2.00	23.00	3.00	25.00	.	3.00	
2	2.00	4.00	2.00	4.00	2.00	3.00	3.00	2.00	22.00	3.00	20.00	.	5.00	
3	5.00	2.00	5.00	2.00	5.00	10.00	2.00	2.00	21.00	3.00	30.00	.	3.00	
4	5.00	5.00	5.00	2.00	2.00	2.00	2.00	2.00	21.00	1.00	20.00	.	2.00	
5	2.00	2.00	3.00	2.00	4.00	2.00	2.00	2.00	21.00	1.00	5.00	.	6.00	
6	1.00	2.00	1.00	2.00	2.00	4.00	3.00	2.00	20.00	1.00	10.00	.	6.00	
7	4.00	4.00	4.00	3.00	4.00	3.00	2.00	2.00	21.00	1.00	20.00	.	5.00	
8	3.00	5.00	3.00	2.00	4.00	2.00	2.00	2.00	25.00	1.00	35.00	.	6.00	
9	5.00	4.00	4.00	4.00	2.00	6.00	6.00	2.00	21.00	1.00	20.00	.	3.00	
10	4.00	3.00	4.00	5.00	3.00	2.00	2.00	1.00	21.00	1.00	30.00	.	6.00	
11	3.00	2.00	2.00	4.00	2.00	15.00	6.00	2.00	21.00	1.00	25.00	.	4.00	
12	4.00	4.00	4.00	4.00	3.00	.00	.00	2.00	21.00	1.00	25.00	.	2.00	
13	4.00	4.00	4.00	4.00	4.00	30.00	12.00	1.00	21.00	1.00	20.00	.	3.00	
14	4.00	3.00	4.00	4.00	4.00	20.00	2.00	2.00	21.00	1.00	20.00	.	4.00	
15	5.00	4.00	4.00	3.00	4.00	100.00	4.00	2.00	21.00	1.00	35.00	.	6.00	
16	2.00	5.00	3.00	4.00	4.00	50.00	3.00	2.00	21.00	1.00	25.00	.	3.00	
17	4.00	3.00	4.00	4.00	4.00	20.00	3.00	1.00	22.00	1.00	20.00	.	4.00	
18	4.00	2.00	4.00	2.00	4.00	10.00	7.00	2.00	21.00	3.00	25.00	.	5.00	
19	3.00	5.00	3.00	5.00	3.00	5.00	4.00	1.00	25.00	1.00	25.00	.	5.00	

[그림 13-8] 역코딩을 통한 새 변수의 정의 2

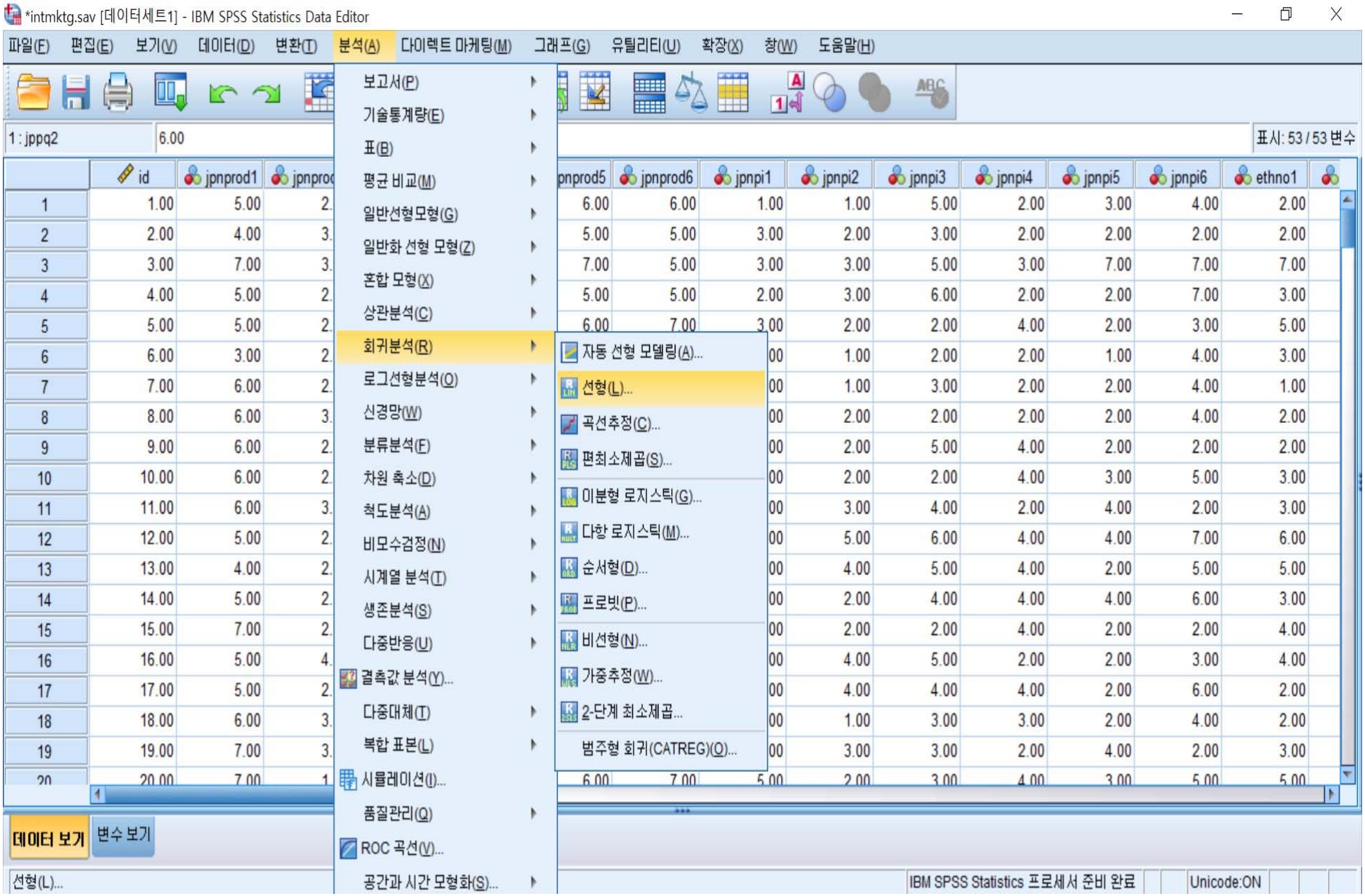


[그림 13-9] 역코딩을 통한 새 변수의 생성 2



	b4	mc4	sn1	sn2	jpnpur1	jpnpur2	gender	age	job	mexpense	var00001	jpi3	jppq2	변수	변수
1	2.00	2.00	2.00	2.00	15.00	4.00	2.00	23.00	3.00	25.00	.	3.00	6.00		
2	4.00	2.00	4.00	2.00	3.00	3.00	2.00	22.00	3.00	20.00	.	5.00	5.00		
3	2.00	5.00	2.00	5.00	10.00	2.00	2.00	21.00	3.00	30.00	.	3.00	5.00		
4	5.00	5.00	2.00	2.00	2.00	2.00	2.00	21.00	1.00	20.00	.	2.00	6.00		
5	2.00	3.00	2.00	4.00	2.00	2.00	2.00	21.00	1.00	5.00	.	6.00	6.00		
6	2.00	1.00	2.00	2.00	4.00	3.00	2.00	20.00	1.00	10.00	.	6.00	6.00		
7	4.00	4.00	3.00	4.00	3.00	2.00	2.00	21.00	1.00	20.00	.	5.00	6.00		
8	5.00	3.00	2.00	4.00	2.00	2.00	2.00	25.00	1.00	35.00	.	6.00	5.00		
9	4.00	4.00	4.00	2.00	6.00	6.00	2.00	21.00	1.00	20.00	.	3.00	6.00		
10	3.00	4.00	5.00	3.00	2.00	2.00	1.00	21.00	1.00	30.00	.	6.00	6.00		
11	2.00	2.00	4.00	2.00	15.00	6.00	2.00	21.00	1.00	25.00	.	4.00	5.00		
12	4.00	4.00	4.00	3.00	.00	.00	2.00	21.00	1.00	25.00	.	2.00	6.00		
13	4.00	4.00	4.00	4.00	30.00	12.00	1.00	21.00	1.00	20.00	.	3.00	6.00		
14	3.00	4.00	4.00	4.00	20.00	2.00	2.00	21.00	1.00	20.00	.	4.00	6.00		
15	4.00	4.00	3.00	4.00	100.00	4.00	2.00	21.00	1.00	35.00	.	6.00	6.00		
16	5.00	3.00	4.00	4.00	50.00	3.00	2.00	21.00	1.00	25.00	.	3.00	4.00		
17	3.00	4.00	4.00	4.00	20.00	3.00	1.00	22.00	1.00	20.00	.	4.00	6.00		
18	2.00	4.00	2.00	4.00	10.00	7.00	2.00	21.00	3.00	25.00	.	5.00	5.00		
19	5.00	3.00	5.00	3.00	5.00	4.00	1.00	25.00	1.00	25.00	.	5.00	5.00		

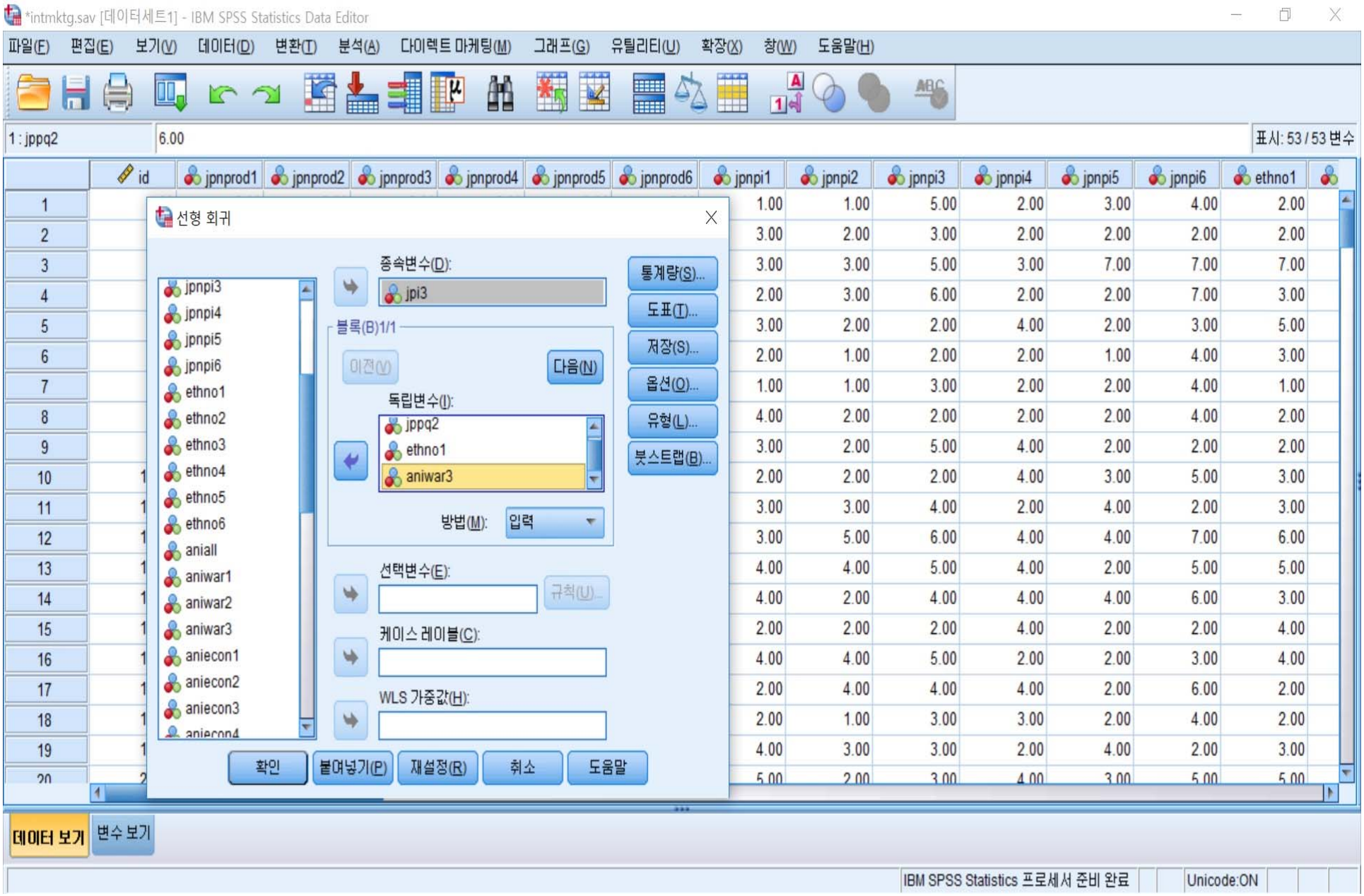
[그림 13-10] 선형회귀분석의 실행



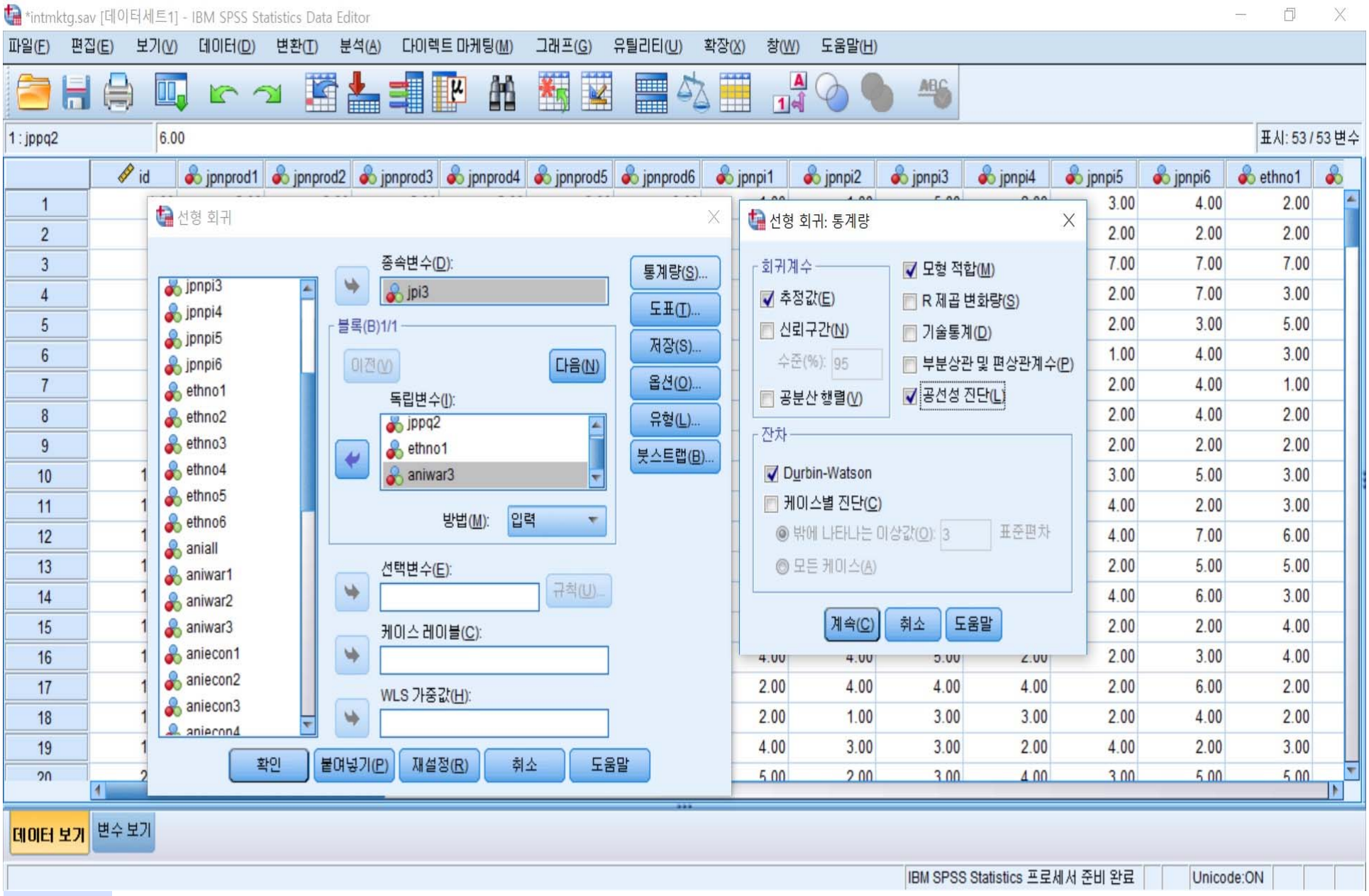
- [그림 13-10]과 같이 **분석(A)** → **회귀분석(R)** → **선형(L)**을 차례로 선택하게 되면 [그림 13-11]과 같은 '선형회귀분석' 창이 제시됨
- 이 창에서는 우선 '종속변수(D)'와 '독립변수(I)'를 지정할 수 있음
- 다음, 독립변수를 회귀식에 투입하는 구체적인 '방법(M)'을 선택

  예) [그림 13-11]에서는 '입력'이란 방법이 기본으로 지정
- '독립변수(I)'의 지정시에 하나의 독립변수만을 선택하고 회귀분석을 진행한다면 단순회귀분석을 수행하는 것이 될 것이고, 2개 이상의 독립변수를 지정하고 회귀분석을 진행한다면 다중회귀분석을 수행하는 것이 될 것임
- '선택변수(E)'란 표본 중의 일부만을 대상으로 선형회귀분석을 실행하기를 원하는 경우에 그 일부를 선택하는 데 사용할 변수를 지정하는 것임

[그림 13-11] 선형회귀분석 창



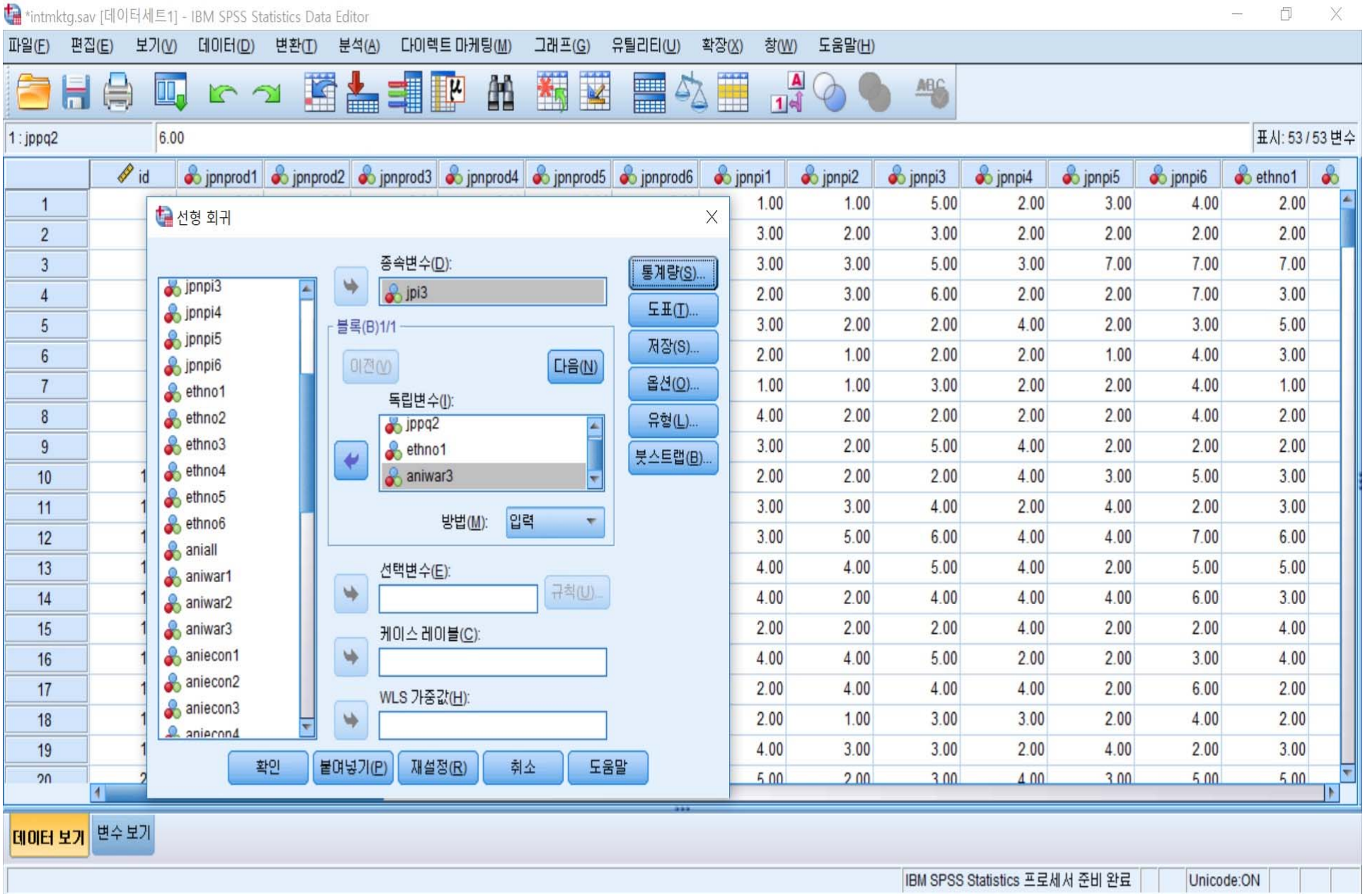
[그림 13-12] 선형회귀분석시 통계량 창



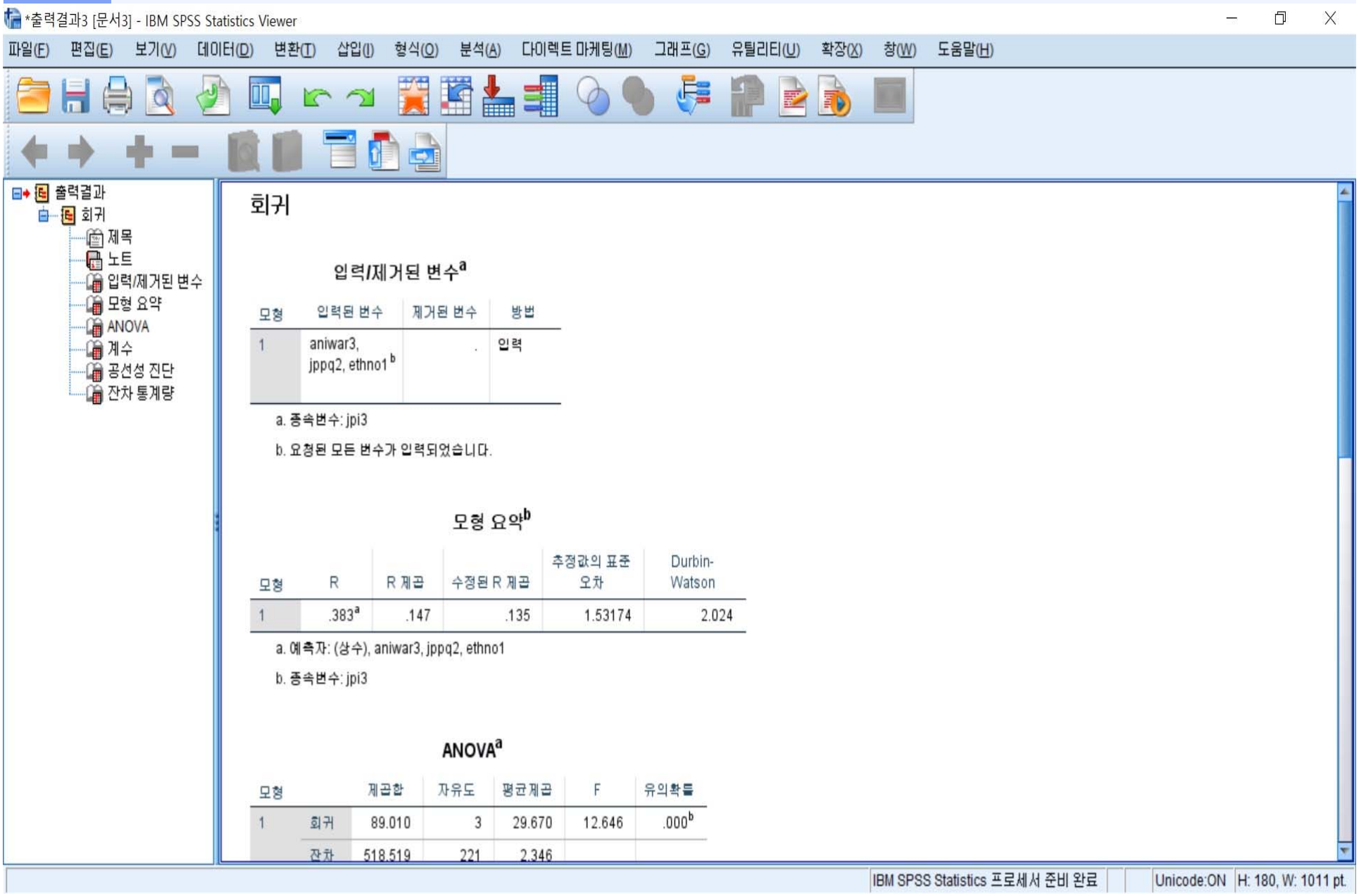
<표 13-6> 통계량 창에서의 선택사항

회귀계수	추정값(E)	회귀계수 추정값(회귀계수, 베타, 표준오차, *t*-value, 유의수준)을 표시해 줍니다.
	신뢰구간(N)	회귀계수에 대한 신뢰구간을 표시해 줍니다.
	공분산행렬(V)	비표준화 회귀계수에 대한 분산-공분산 행렬을 표시해 줍니다. 대각선에는 분산이 표시되고 대각선의 위와 아래에는 공분산이 표시됩니다.
모형적합(M)		R^2, 다중 R, 수정된 R^2, 표준오차 등을 표시해 주고, 분산분석표에서는 자유도, 제곱합, 제곱평균, F값 등을 표시해 줍니다.
R제곱 변화량(S)		독립변수를 추가하거나 삭제하는데 따르는 R^2의 변화정도를 표시해 줍니다.
기술통계(D)		각 변수의 평균, 표준편차, 각 변수들 간의 상관관계를 표시됩니다.
공선성진단(L)		개별 독립변수에 대한 공차한계와 다중공선성 진단을 위한 기타 통계량을 표시해 줍니다.

[그림 13-13] 선형회귀분석의 실행



[그림 13-14] 선형회귀분석 결과 창



* 최근 수십 년 간 logit 회귀분석, Poisson 회귀분석, 구조방정식(structural equation modeling) 등과 같은 회귀분석의 용도와 유사한 세련된 분석방법 들이 출현했으나, 분석개념의 이해와 사용의 용이성 측면 등에서 아직도 (선형) 회귀분석이 많은 사랑을 받고 있다고 할 수 있음

1. 회귀분석과 상관관계분석

▫ 단순회귀선과 상관계수는 동일한 현상을 조금 다른 방법으로 설명해 줌

♦ **상관계수**

▫ 단순회귀선을 중심으로 한 측정자료의 산포정도(degree of scatter)를 측정

♦ **회귀계수**

▫ 단순회귀선의 기울기(slope)를 측정

♦ 단순회귀선의 기울기의 크기와 자료의 산포정도간에는 관계가 거의 없음

단순회귀분석의 회귀계수와 상관계수 (식 13-13)

회귀계수 = COV_{XY} / S_X^2 (a)

상관계수 = $COV_{XY} / S_X S_Y$ (b)

- **상관계수와 단순회귀분석간의 관계**
  - 상관계수의 제곱(=r^2)이 바로 단순회귀식의 설명력(=R^2)과 동일
  - 회귀계수나 상관계수 모두 분석대상 현상(즉, X와 Y)간의 관계가 선형(linear)이라고 하는 가정 *(실제 변수간 관계가 비선형이면 둘 다 모두 상당한 오류의 가능성을 내포하게 됨)*
  - 단순회귀분석의 경우, 종속변수와 독립변수를 서로 바꾸면(물론 인과관계가 이처럼 쉽게 바뀔 수는 없지만), 2개의 회귀선이 계산될 것이고 일반적으로 이 두 개의 회귀선은 동일하지 않음
    - ☞ Y를 종속변수로 하는 회귀계수는 COV_{XY}/S_X^2 가 될 것이고, X를 종속변수로 하는 회귀계수는 COV_{XY}/S_Y^2 가 되기 때문임

2. 근원척도와 회귀분석

-> 사회현상이 측정된 근원척도의 유형에 따라 적용 가능한 회귀분석의 유형이 달라질 수 있음

1] 등간/비율척도 독립변수와 등간/비율척도 종속변수

- 전형적인 회귀분석 가능

2] 명목척도 독립변수와 등간/비율척도 종속변수

- 추가적인 준비(예: dummy coding)를 통해 회귀분석 가능

3) 등간/비율척도 독립변수와 명목척도 종속변수

▫ logit모형이나 probit모형으로 분석가능

☞ logit모형(logit model 혹은 logistic regression)은 probit모형에 비해 상대적으로 계산이 용이한 장점을 가지고 있음

♦ **logit**

▫ 로지스틱회귀분석의 종속변수가 1이 될 확률이 logistic probability unit이라고 표현되는 데서 유래한 용어

♦ 명목척도 수준으로 측정된 독립변수와 명목척도 수준으로 측정된 종속변수(=특정 집단에 속할 확률)

▫ 가중 최소자승 logit모형이 유용하게 사용될 수 있음

☞ 명목척도란 사회현상에 대한 정보의 종류 및 양이 가장 적은 종류의 근원척도이므로 불가피한 경우가 아니라면 독립변수, 종속변수를 명목척도의 수준으로 측정하지 않는 것이 사회현상간의 관계를 정확히 분석하는 데 유리함

3. 회귀분석과 판별분석

♦ **판별분석**(discriminant analysis)

▫ 로지스틱 회귀분석과 유사하게 등간척도/비율척도 수준으로 측정된 독립변수와 명목척도 수준으로 측정된 종속변수간의 관계를 분석해 주는 통계기법

▫ 특정 집단에의 소속(group membership)을 결정하는 변수가 무엇인지, 그 영향을 어느 정도인지를 분석하는 기능을 수행하므로, 세분시장 예측, 부도기업 예측 등 다양한 분야에서 사용

♦ **로지스틱회귀분석과 판별분석의 차이점**

① **판별분석**에는 독립변수들의 정규성(normality)과 각 집단의 분산/공분상의 동일성 등 엄격한 가정이 성립되어야 하나, **로지스틱회귀분석**에는 이러한 엄격한 가정이 불필요해서 현실적인 사용가능성이 높다는 점

② **판별분석**에서는 판별점수를 계산하여 소속집단을 예측하는 데 비해, **로지스틱회귀분석**에서는 측정값이 특정집단에 속할 확률을 기초로 소속집단을 예측한다는 점