18장. 로지스틱 회귀분석 **

18.1 로지스틱 회귀분석의 개요

로지스틱 회귀분석의 사용

로지스틱 회귀분석(logistic regression)은 종속변수가 명목변수일 때 사용하는 회귀분석 방법이다. 회귀분석과 모든 형태가 같고 단지 종속변수만 이항형 또는 순서적인 다항형인 경우에 사용한다.

종속변수가 이항형(dichotomous)일 때 일반적인 선형회귀모형에 의한 분석을 적용할 수 없는 이유를 설명하고자 한다.

예를 들어 종속변수가 범주형으로 0, 1의 값을 갖는 경우에 우리는 다음과 같은 모형을 가정할 수 있다. 즉, 측정 불가능한 잠재변수의 값에 따라 측정 가능한 실현값은 0 또는 1을 갖는 다음과 같은 절편이 없는 선형 확률모형으로 표현한다.

$$y_j = \beta x_j + \epsilon_j ; j = 1,...,n,$$

여기서 $y_i = \begin{cases} 1, & y^* > 0 \\ 0, & y^* \le 0 \end{cases}$, 그리고 y^* 는 잠재변수(latent variable)이다.

일반적인 회귀모형에서는 $\epsilon_i \sim N(0, \sigma^2)\ (i = 1,2,...,n)$와 같이 모든 오차항의 분산이 같다는 등분산 가정이 있다. 그런데 위 모형에서는 이러한 등분산 가정이 성립하지 않는다.[1] 즉, 분산이 x_i에 종속되어 등분산 가

---

1) 이를 증명하기 위해 식을 오차항을 중심으로 다시 표현하면 다음과 같다.

$$\epsilon_i = \begin{cases} 1 - \beta x_i, & y_i = 1 \\ -\beta x_i, & y_i = 0. \end{cases}$$

그리고 이산형 확률변수의 기댓값 공식에 의하면

$$E(y_i) = \sum y_i f(y_i) = 0 \times \Pr(y_i = 0) + 1 \times \Pr(y_i = 1).$$

그러므로

$$\Pr(y_i = 1) = E(y_i) = \beta x_i, \text{ 그리고 } \Pr(y_i = 0) = 1 - \beta x_i.$$

정이 성립하지 않는다. 그러므로 일반적인 최소제곱법(OLS: ordinary least squares)이 아닌 가중최소제곱법(WLS: weighted least squares)으로 회귀계수 β를 추정하여야 한다.

그러나 WLS는 다음과 같은 문제점이 있다. WLS를 위한 가중값은 오차항의 분산인데 이 값을 알 수 없으므로 다음과 같은 추정값을 사용한다.

$$w_i = \widehat{Var}(\epsilon_i) = \hat{\beta} x_i (1 - \hat{\beta} x_i),$$

여기서 $\hat{\beta}$는 OLS에 의한 추정값이다.

그런데 이 가중값을 사용하는 데는 세 가지 문제점이 있다.

첫 째, w_i가 음수일 가능성이 존재한다.

둘 째, ϵ_i가 정규분포를 따르지 않는다.

셋 째, $\hat{\beta} x_i = \widehat{\Pr}(Y_i = 1)$이 (0,1)의 범위가 아닐 수 있다.

그러므로 WLS도 이에 대한 해법이 될 수 없다. 그래서 종속변수가 이항형일 때 일반적인 회귀분석으로는 해법을 구할 수 없고 새로운 방법이 필요하다.

다음 절은 종속변수가 0 또는 1을 갖는 베르누이 분포를 따른다는 사실을 이용하여 종속변수가 1을 가질 확률의 승산비 등을 추정하는 방식으로 로지스틱회귀분석을 소개한다.

모형의 오차항, ϵ_i의 분산은 다음과 같이 유도할 수 있다.

$$\begin{aligned} Var(\epsilon_i) &= E(\epsilon_i^2) - E(\epsilon_i)^2 \;= \sum \epsilon_i^2 f(\epsilon_i) - 0 \\ &= (1-\beta x_i)^2 \Pr(y_i = 1) + \beta^2 x_i^2 \Pr(y_i = 0) \\ &= (1-\beta x_i)^2 \beta x_i + \beta^2 x_i^2 (1 - \beta x_i) \\ &= \beta x_i (1 - \beta x_i). \quad (\text{즉, 분산이 } x_i \text{에 종속한다}) \end{aligned}$$

18.2 로지스틱 회귀분석의 절차

로지스틱 회귀분석 모형의 유도

로지스틱 회귀분석모형의 원리를 알아야 계수의 의미를 파악하고 모형의 해석이 가능하다.

앞의 회귀모형을 다시 일반화하여 적으면 (편의상 절편은 0이라 가정) 잠재변수 y_i*는 관측할 수 없지만 다음 식을 따른다고 가정할 수 있다.

$$y_i^* = \sum_{j=1}^{k} \beta_j x_{ij} + e_i = z_i + e_i;\ i = 1,...,n,$$

실제로 관측할 수 있는 y_i는 다음과 같이 적을 수 있다.

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & y_i^* \le 0. \end{cases}$$

그러므로 $\Pr(y_i = 1) = \Pr(e_i > -\sum \beta_j x_{ij}) = 1 - F(-\sum \beta_j x_{ij})$.

이때, 오차항 e_i가 0에 대하여 좌우대칭이고 분포함수 F를 갖는다고 가정하면, [2] $\Pr(y_i = 1)$는 다음과 같이 표현할 수 있다.

$$p_i \equiv \Pr(y_i = 1) = F(\sum \beta_j x_{ij}) \equiv F(z_i).$$

모형에서 e_i의 분포를 정규분포라 가정하는 것(즉, $F \equiv$ Normal)이 일반적일 것이다. 그러나 정규분포일 때보다 로지스틱분포일 때 훨씬 계산식이 간단한 형태로 표현될 수 있고 두 분포의 차이가 크지 않기 때문에 [3] e_i의 분포를 로지스틱분포로 가정하도록 하겠다. [4]

---

2) 확률변수 X의 분포함수는 $F(x) = \Pr(X \le x)$. X가 0에 대하여 대칭이라면 $\Pr(X > x) = 1 - \Pr(X \le -x) = 1 - F(-x)$.

3) 표준정규분포함수는 다음과 같다. $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$.

그러므로 F를 다음과 같은 logistic 분포함수로 근사 대체한다.

$$F(x) = \frac{e^x}{1+e^x} \text{ , 또는 } x = \ln\frac{F(x)}{1-F(x)}.$$

$\Pr(y_i=1)$는 다음과 같이 정리할 수 있다.

$$p_i = F(z_i), \Rightarrow z_i = \ln\frac{p_i}{1-p_i}.$$

그러므로 로지스틱회귀모형을 궁극적으로 다음과 같이 표현하게 된다.

$$\ln\frac{p_i}{1-p_i} = \sum\beta_j x_{ij} \text{ , 또는 } \frac{p_i}{1-p_i} = e^{\sum\beta_j x_{ij}} = \exp(\sum\beta_j x_{ij}),$$

여기서 $\ln\frac{p_i}{1-p_i} = \text{logit}(p_i)$를 로짓(logit)이라 하고, $\frac{p_i}{1-p_i}$를 승산(오즈, odds)이라 한다.

즉, 정리하면 실제값이 1이 나올 확률의 로짓이 일반 중회귀 모형을 하고 있다. 또는 확률의 승산이 지수회귀모형을 따른다.

표준로지스틱분포함수는 다음과 같다. $F(x) = \frac{1}{1+\exp(-\pi x/\sqrt{3})}$.

두 분포함수식의 모양은 다르지만 그림으로 그려보면 두 함수는 크게 차이가 나지 않는다. 정규분포함수 $\Phi(x)$를 rescale 하면 두 분포함수의 최대차이는 0.001을 넘지 않는다.

즉, $\max\left|\Phi(\frac{16}{15}x) - F(x)\right| \le 0.001$. 또한 rescale하지 않더라도 최대차이가 0.022를 넘지 않는다.

4) 오차항의 분포를 정규분포로 가정하고 모형을 설정한 것을 프로빗 모형이라 한다.

$$\text{probit model} : \Phi^{-1}(p_j) = \sum\beta_i x_{ij}$$

<주의> F를 정규분포로 가정한 다음 프로빗(probit) 모형과 로지스틱으로 가정한 로짓(logit) 모형은 척도의 차이가 존재하므로 추정 후 계수의 단순 비교는 어렵다.($\sqrt{3}/\pi$배?)

승산, 승산비의 의미

승산(odds)은 어떤 사건이 발생할 확률과 발생하지 않을 확률의 비율이다. 또한 입력변수가 1단위 증가할 때 늘어나는 승산의 비율을 오즈비(승산비, odds ratio)라 한다.

예를 들어 승산이 2이라면 사건이 발생할 확률이 2/3, 발생하지 않을 확률이 1/3이라는 의미이다. Odds에 관한 회귀모형은 다음과 같이 정리할 수 있다.

$$\begin{aligned}\text{Odds} &= \frac{p_i}{1-p_i}\\ &= \exp(\Sigma \beta_j x_{ij})\\ &= \exp(\beta_1 x_{i1})\exp(\beta_2 x_{i2})\cdots\exp(\beta_k x_{ik}).\end{aligned}$$

x_j가 1단위 증가할 때 늘어나는 승산의 비율인 오즈비(승산비, odds ratio)라 하고 다음과 같이 계산된다.

$$\frac{\exp(\beta_1 x_{i1})\exp(\beta_2 x_{i2})\cdots\exp(\beta_j(x_{ij}+1))\cdots\exp(\beta_k x_{ik})}{\exp(\beta_1 x_{i1})\exp(\beta_2 x_{i2})\cdots\exp(\beta_i x_{ij})\cdots\exp(\beta_k x_{ik})} = \exp(\beta_j).$$

즉, Odds ratio$=\exp(\beta_j)$.

회귀계수가 음이라면 승산비는 1보다 작게 되고 이는 x가 증가함에 따라 승산(odds)은 감소한다는 의미가 된다.

$$\text{즉, } \beta_j < 0 \Rightarrow \exp(\beta_j) < 1 \Rightarrow \text{odds ratio} < 1 \Rightarrow \text{odds decrease.}$$

예를 들어, 농구경기 슛의 성공 승산에 관한 자료에서 골에서의 거리가 x_j(단위: meter)일 때 $\beta_j = -0.4$이라면 이는 거리 x_j가 1미터 증가하면 슛 성공 odds는 $e^{-0.4}$배 만큼 증가, 즉 $e^{0.4} = 1.49$배 감소하는 것이다.

회귀계수의 추정

승산(odds), 또는 승산비를 구하기 위해서는 회귀계수를 추정해야 하고 유의한 회귀계수만을 모형에 포함시켜야 한다.

일반적으로 회귀계수를 추정하기 위해 최대우도 추정법(MLE)을 사용한다. 우도함수는 f가 베르누이 분포임($\because$ y_i가 0 또는 1이므로)을 이용하여 다음과 같이 유도할 수 있다.

$$f(y_1, y_2, ..., y_n | \boldsymbol{x}) = \prod_{i=1}^{n} f(y_i | x_{i1}, x_{i2}, ..., x_{ik})$$
$$= \prod_{i=1}^{n} \left(\frac{e^{\Sigma \beta_j x_j}}{1+e^{\Sigma \beta_j x_j}} \right)_i^{y} \left(\frac{1}{1+e^{\Sigma \beta_j x_j}} \right)^{1-y_i}.$$

그리고 이 식을 최대화하는 β를 반복적으로 근사해서 해를 구할 수 있다. 이렇게 구한 회귀계수의 유의성 가설 ($H_0 : \beta_j = 0$) 검정은 다음과 같이 Wald 통계량 z^2가 자유도가 1인 카이제곱분포를 따른다는 사실을 이용한다.

$$z^2 = \left(\frac{\hat{\beta}_j}{Se(\hat{\beta}_j)} \right)^2.$$

그러나 여기서도 일반적 회귀분석과 같이 유의한 입력변수를 모두 회귀모형에 포함시키지 않고 여러 가지 변수선택방법(전진선택, 후진소거, 단계적 방법 등)이 사용된다. 어떠한 변수를 포함한 모형이 가장 우수한가를 판단하는 기준으로 결정계수 외에 AIC(Akaike Information Criterion)를 참조한다.

【참고】 AIC(Akaike Information Criterion)

아카이케 정보는 다음 Kullback-Leibler Information($I(g,f)$)의 개념으로부터 유도된다.

$$I(g,f) = E_Y \log \frac{g(Y)}{f(Y)},$$

여기서 $g(Y)$는 실제모형이고 $f(Y)$는 적합된 모형이다. 또한 $I(g,f) \geqq 0$ 이며, 우리가 적합한 모형이 정확히 실제모형과 일치할 때 최소값 0을 갖는다. ($\because \log 1 = 0$) 그런데

$$I(g,f) = E_Y \log g(Y) - E_Y \log f(Y)$$

이고, 이러한 정보지수를 최소화하는 모형이 최적모형이므로 앞의 실제모형 항은 고정되어 있다고 보고 $E_Y \log f(Y)$를 최대화 하여야 한다.

그런데 이 값은 다음과 같이 근사시킬 수 있다.

$$E_Y \log f(Y) \approx \frac{1}{N} \sum \log f(y_n), \ N\text{은 표본수.}$$

그리고 실제로는 윗 식 대신에 $\sum \log f(y_n | \hat{\theta}) / N$를 사용하는데 이 값으로는 편의(bias)가 존재한다. 왜냐하면 y_n를 이용해 θ를 추정하고 다시 $\hat{\theta}$를 가정해 추정하기 때문에 과적합 문제가 발생한다. 그래서 $E_Y \log f(Y)$의 불편추정량을 이용하는데 이러한 과정에서 AIC가 유도된 것이다. 즉,

$$E\left(E_Y \log f(Y | \hat{\theta}) - \frac{1}{N} \sum \log f(y_n | \hat{\theta}) \right) \approx -\frac{k}{N}$$

이므로 다음과 같이 볼 수 있다.

$$E_Y \log f(Y) \approx \frac{1}{N} \sum \log f(y_n | \hat{\theta}) - \frac{k}{N} \ .$$

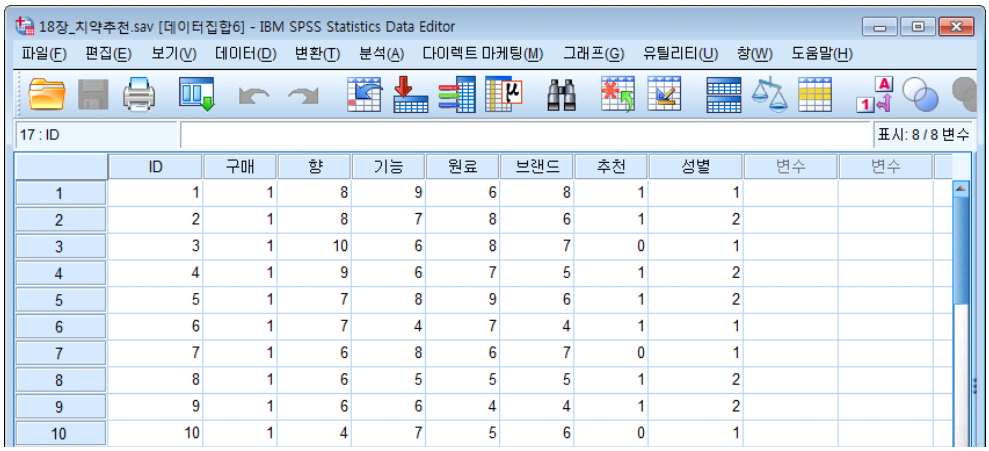
이때 AIC를 다음과 같이 정의한다.

$$AIC = -2 \sum \log f(y_n | \hat{\theta}) + 2k \quad (\approx -2N \cdot E_Y \log f(Y))$$

이는 카이제곱분포를 따른다. 변수추가가 유의하게 모형을 적합 시키는가를 카이제곱 검정으로 판단할 수 있도록 $E_Y \log f(Y)$에 $-2N$ 을 곱하여 준 것이다. 결론적으로 AIC를 최소화하는 모형이 최적의 모형이 된다.

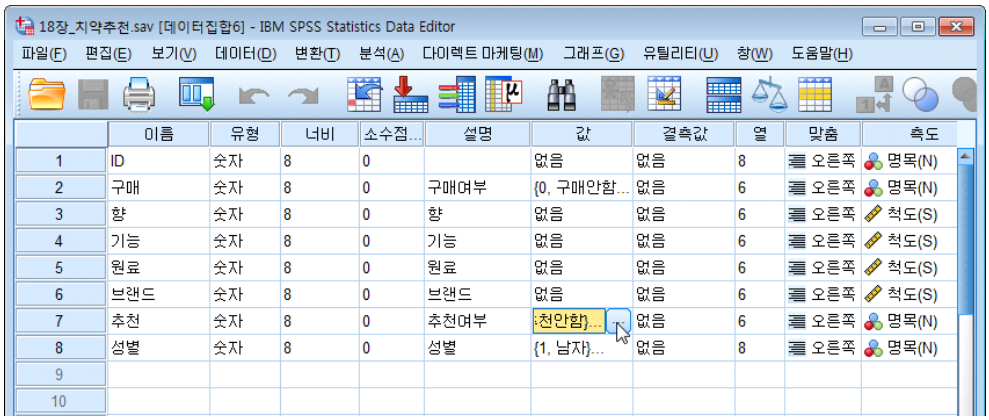
【예제 18.1】 앞 장의 【예제 17.1】에서 치약의 속성에 대하여 만족도를 평가하고 재구매 여부를 물었는데 추가적으로 타인에게 추천할 것인지를 물었다. 각 속성에 대한 만족도가 타인에게 추천여부에 어떤 영향을 주는지, 또 만족도가 상승할 때 추천확률이 어느 정도 증가하는지를 분석하시오. (자료명: "18장_치약추천.sav")

<풀이> 다음과 같이 자료를 입력하였다.



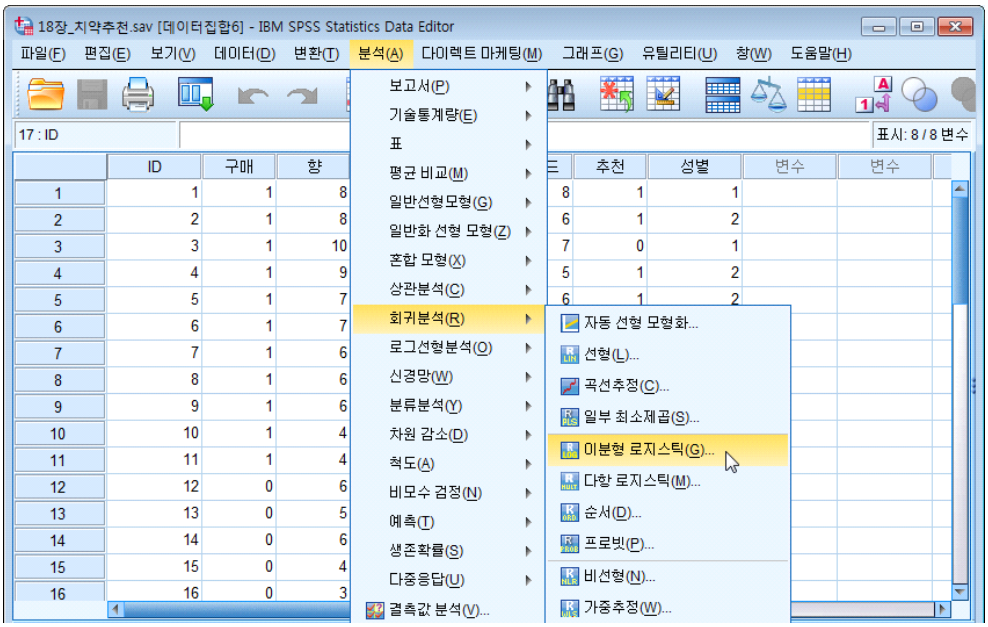
	ID	구매	향	기능	원료	브랜드	추천	성별	변수	변수
1	1	1	8	9	6	8	1	1		
2	2	1	8	7	8	6	1	2		
3	3	1	10	6	8	7	0	1		
4	4	1	9	6	7	5	1	2		
5	5	1	7	8	9	6	1	2		
6	6	1	7	4	7	4	1	1		
7	7	1	6	8	6	7	0	1		
8	8	1	6	5	5	5	1	2		
9	9	1	6	6	4	4	1	2		
10	10	1	4	7	5	6	0	1		

변수들은 다음과 같으며 여기서는 '추천' 변수가 종속변수, '향', '기능', '원료', '브랜드' 등(계량변수)과 함께 '성별'(명목 변수)이 독립변수가 된다.



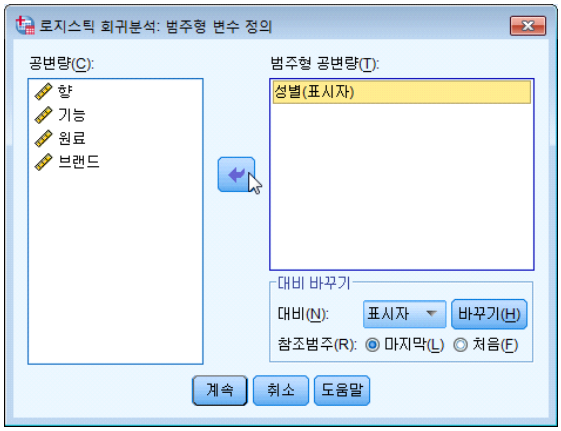
	이름	유형	너비	소수점...	설명	값	결측값	열	맞춤	측도
1	ID	숫자	8	0		없음	없음	8	오른쪽	명목(N)
2	구매	숫자	8	0	구매여부	{0, 구매안함...	없음	6	오른쪽	명목(N)
3	향	숫자	8	0	향	없음	없음	6	오른쪽	척도(S)
4	기능	숫자	8	0	기능	없음	없음	6	오른쪽	척도(S)
5	원료	숫자	8	0	원료	없음	없음	6	오른쪽	척도(S)
6	브랜드	숫자	8	0	브랜드	없음	없음	6	오른쪽	척도(S)
7	추천	숫자	8	0	추천여부	{1, 천안함}...	없음	6	오른쪽	명목(N)
8	성별	숫자	8	0	성별	{1, 남자}...	없음	8	오른쪽	명목(N)
9										
10										

[분석] 메뉴에서 <회귀분석>-<이분형 로지스틱>을 선택한다.

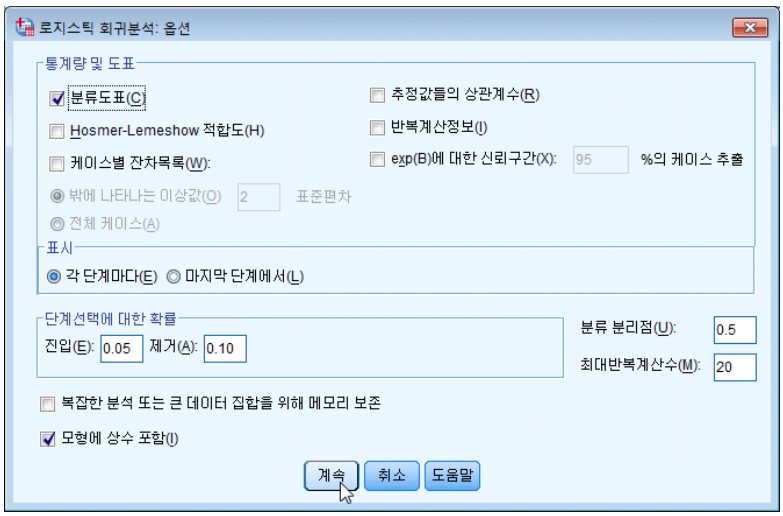


<종속변수> 창에 '추천'을 입력하고 독립변수들은 <공변량> 창에 입력한다. 이때 공변량은 보통 계량형 변수이므로 이들 중 명목 또는 순서형 변수인 경우에는 [범주형] 버튼을 눌러 지정해주어야 한다.

[범주형 변수 정의] 대화창에서 <범주형 공변량>에 '성별' 변수를 지정해 준다.



다시 [로지스틱 회귀모형] 대화창으로 돌아가 [옵션] 버튼을 눌러 [로지스틱 회귀분석: 옵션] 대화상자로 이동한다. 여기서는 분류의 정확도를 표시하는 '□분류도표'에 체크한다.



다시 [로지스틱 회귀모형] 대화창에서 변수선택 [방법]을 선택하는데 여기서는 전진선택 방법인 [앞으로: Wald]를 클릭한다.

<출력결과>

케이스 처리 요약

가중되지 않은 케이스[a]		N	퍼센트
선택 케이스	분석에 포함	16	100.0
	결측 케이스	0	.0
	합계	16	100.0
비선택 케이스		0	.0
합계		16	100.0

a. 가중값을 사용하는 경우에는 전체 케이스 수의 분류표를 참조하십시오.

종속변수 코딩

원래 값	내부 값
추천안함	0
추천함	1

범주형 변수 코딩

		빈도	파라미터 코딩 (1)
성별	남자	9	1.000
	여자	7	.000

자료의 요약을 보면 분석하는 자료는 모두 16개 이다.

종속변수 코딩은 원자료와 동일하게 '추천안함'이 0, '추천함'이 1이다. 즉, '추천안함'이 기준값 0이고 '추천'하는 것이 1이므로 출력결과에 사용되는 확률은 추천할 확률이다.

또한 범주형변수 코딩에서는 원자료는 '남자'=1, '여자'=2로 코딩되었지만 파라미터 코딩을 자동으로 여자를 기준값 0으로 하고 남자에 1을 부여하였다. 이는 여자가 0, 남자가 1이므로 오즈를 해석할 때 "여자에 비해 남자는..." 하고 해석하여야 한다.

분류표[a]

관시됨			예측		
			추천여부		
			추천안함	추천함	분류정확 %
1 단계	추천여부	추천안함	7	1	87.5
		추천함	2	6	75.0
	전체 퍼센트				81.3
2 단계	추천여부	추천안함	7	1	87.5
		추천함	2	6	75.0
	전체 퍼센트				81.3

a. 절단값은 .500입니다.

방정식에 포함된 변수

		B	S.E,	Wals	자유도	유의확률	Exp(B)
1 단계[a]	성별(1)	-3.045	1.345	5.122	1	.024	.048
	상수항	1.792	1.080	2.752	1	.097	6.000
2 단계[b]	향	.854	.472	3.270	1	.071	2.349
	성별(1)	-3.964	1.915	4.284	1	.038	.019
	상수항	-2.989	2.689	1.236	1	.266	.050

a. 변수가 1: 단계에 진입했습니다 성별. 성별.

b. 변수가 2: 단계에 진입했습니다 향. 향.

방정식에 포함되지 않은 변수

			점수	자유도	유의확률
1 단계	변수	향	4.842	1	.028
		기능	.780	1	.377
		원료	.854	1	.355
		브랜드	1.145	1	.285
	전체 통계량		5.483	4	.241
2 단계	변수	기능	.048	1	.827
		원료	.273	1	.601
		브랜드	.085	1	.770
	전체 통계량		.936	3	.817

출력결과에 의하면 2단계에 걸쳐 변수가 선택되는 데 처음에는 '성별' 변수가 다음으로는 '향' 변수가 선택되었다.

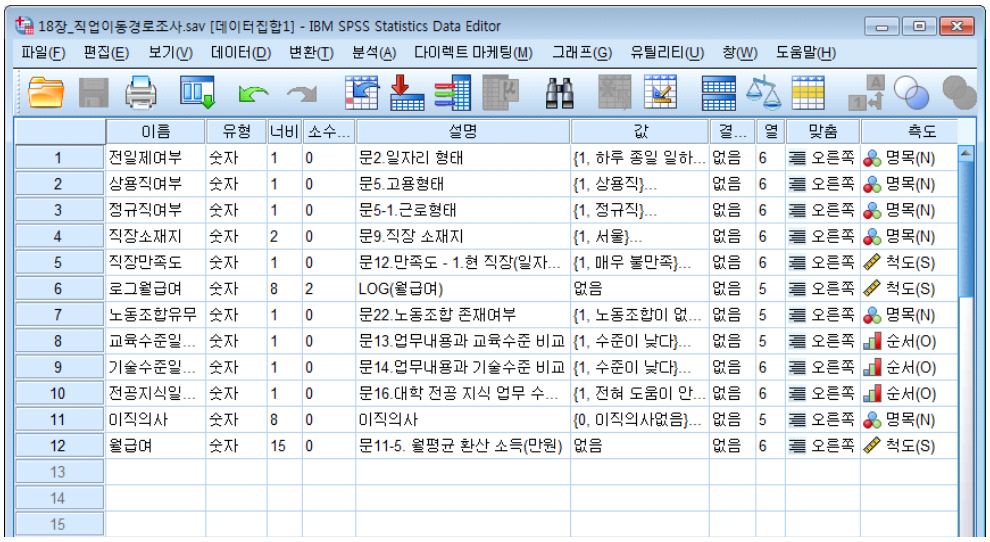
함수에 의한 분류 정확도는 81.3%로, 16명 중 13명을 정확히 예측하였다. 진입한 두 변수, '향'과 '성별'이 '추천'에 유의한 영향을 주고 있는데 각각의 유의확률은 0.071, 0.038 등이다. 계수를 해석하면 '향'에 대한 만족도가 1점 증가하면 추천승산이 2.349배 증가하고 여자에 비해 남자는 추천 승산이 0.19배로 줄어든다. (회귀계수는 음수)

'향'에 대한 유의확률이 0.05보다 낮지 않은데 모형에 포함된 이유는 1단계에서 유의확률이 0.028(방정식에 포함되지 않은 변수 출력결과)이어서 모형에 진입했고 모형의 다른 변수와 함께 유의확률이 0.071로 0.1보다 높지는 않아 제거되지 않은 것이다. ■

【예제 18.2】 다음은 노동부와 한국고용정보원에서는 2006년 고학력 청년층의 취업실태와 취업 이후 노동시장에서의 이동을 지속적으로 조사·분석하기 위하여 전문대 및 4년제 대학 졸업자들을 대상으로 조사한 「대졸자 직업이동경로조사」 자료의 일부분이다. ($n=20{,}244$)

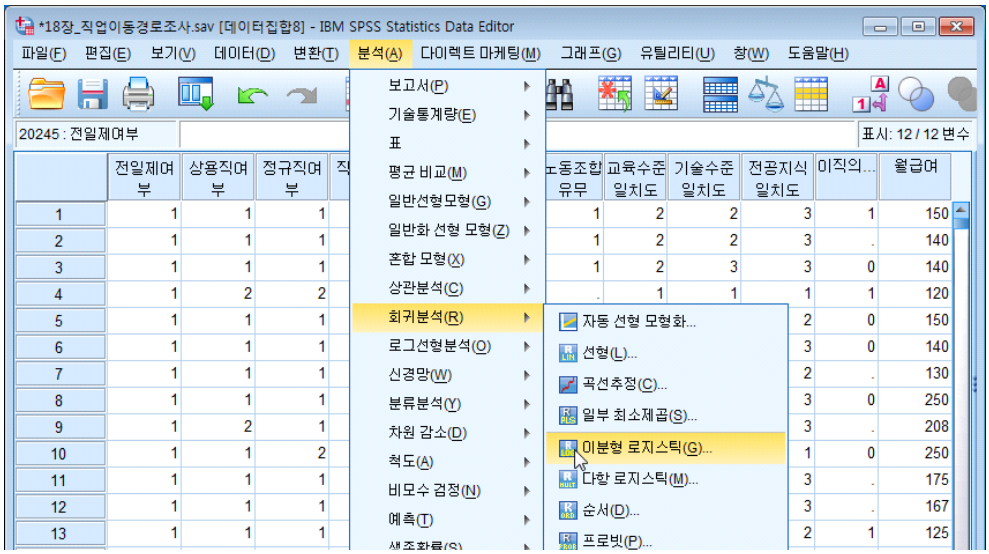
직장인들의 이직의사(y)에 영향을 줄 수 있는 독립변인들로는 전일제여부(X_1), 상용직여부(X_2), 정규직여부(X_3), 직장소재지(X_4), 직무만족도(X_5), 로그월급여(X_6), 노조유무(X_7), 교육수준일치도(X_8), 기술수준 일치도(X_9), 전공 일치도(X_{10}) 등 10개 변수를 선택하였고 여기서는 $X_1 - X_7$의 7개 변수에 대하여만 분석하고자 한다. (자료명: "18장_직업이동경로조사.sav")

<풀이> 자료는 다음과 같은 형태로 입력이 된다. [변수보기] 시트에서 사용된 변수들을 살펴보면 종속변수는 '이직의사'이고 여기에 영향을 미치는 직장의 고용 형태, 급여, 근무 형태, 만족도, 전공일치(분석 제외) 등에 관련된 독립변수로 분류될 수 있다. 종속변수는 이분형이고 독립변수에 계량변수와 명목변수가 혼합되어 있음을 알 수 있으므로 이분형 로지스틱회귀분석으로 분석해야 한다,

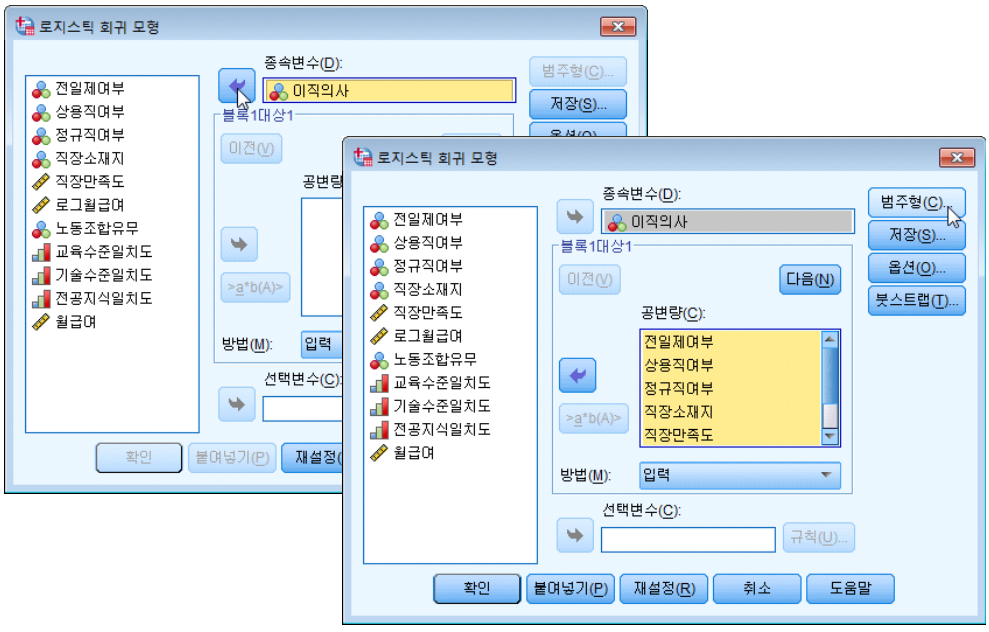


	이름	유형	너비	소수...	설명	값	결...	열	맞춤	측도
1	전일제여부	숫자	1	0	문2.일자리 형태	{1, 하루 종일 일하...	없음	6	오른쪽	명목(N)
2	상용직여부	숫자	1	0	문5.고용형태	{1, 상용직}...	없음	6	오른쪽	명목(N)
3	정규직여부	숫자	1	0	문5-1.근로형태	{1, 정규직}...	없음	6	오른쪽	명목(N)
4	직장소재지	숫자	2	0	문9.직장 소재지	{1, 서울}...	없음	6	오른쪽	명목(N)
5	직장만족도	숫자	1	0	문12.만족도 - 1.현 직장(일자...	{1, 매우 불만족}...	없음	6	오른쪽	척도(S)
6	로그월급여	숫자	8	2	LOG(월급여)	없음	없음	5	오른쪽	척도(S)
7	노동조합유무	숫자	1	0	문22.노동조합 존재여부	{1, 노동조합이 없...	없음	5	오른쪽	명목(N)
8	교육수준일...	숫자	1	0	문13.업무내용과 교육수준 비교	{1, 수준이 낮다}...	없음	5	오른쪽	순서(O)
9	기술수준일...	숫자	1	0	문14.업무내용과 기술수준 비교	{1, 수준이 낮다}...	없음	6	오른쪽	순서(O)
10	전공지식일...	숫자	1	0	문16.대학 전공 지식 업무 수...	{1, 전혀 도움이 안...	없음	6	오른쪽	순서(O)
11	이직의사	숫자	8	0	이직의사	{0, 이직의사없음}...	없음	5	오른쪽	명목(N)
12	월급여	숫자	15	0	문11-5. 월평균 환산 소득(만원)	없음	없음	6	오른쪽	척도(S)
13										
14										
15										

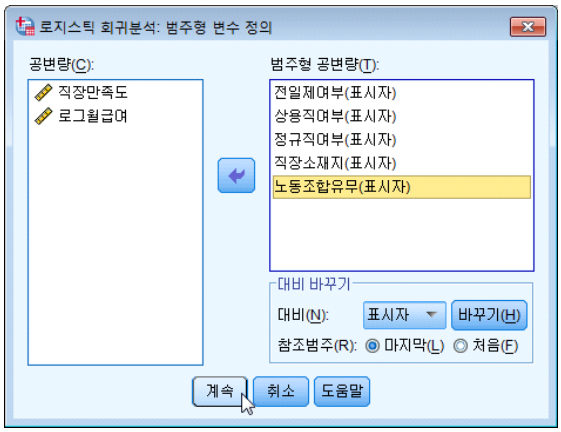
[분석] 메뉴의 <회귀분석>-<이분형 로지스틱>을 선택한다.



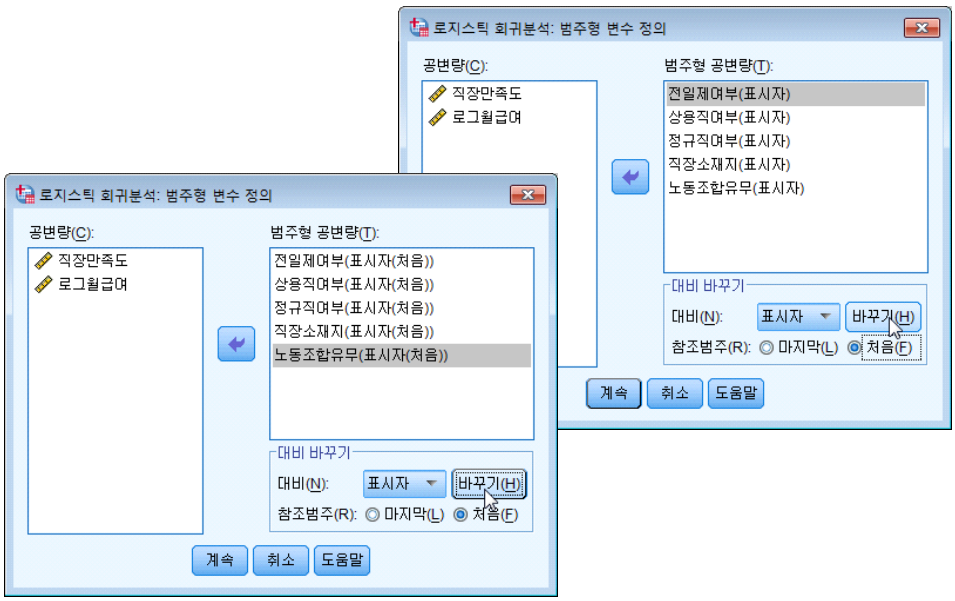
[로지스틱 회귀모형] 대화상자에서 <종속변수>로는 '이직의사' 변수를 입력하고 <공변량>에는 독립변수인 7개 변수를 지정 입력하고 범주형 변수를 정의한다.



[범주형 변수 정의]에서 <공변량> 변수 중에 범주형인 변수를 지정한다.



범주형 변수들은 사후에 오즈비(한 단위 증가시 오즈의 변화율) 해석을 위하여 변수값의 기준이 되는 항목(예. 서울)을 지정해주는 작업이 필요하다. 기준(참조범주) 값이 0, 다른 항목이 1의 값을 갖게 하는 <표시자> '대비'에서 '참조범주'를 <처음>으로 정하고 <바꾸기>를 각 변수마다 클릭하여 변경해준다.



[옵션]에서는 <분류표>와 <적합도>에 체크하도록 한다.

[로지스틱 회귀 모형] 대화상자에서 변수선택 방법으로 전진선택방법인 [앞으로: Wald]를 선택하고 [확인]을 눌러 실행한다.

<출력결과> 모두 20,244케이스 중에서 이직에 관련된 의사를 밝힌 12,657명을 대상으로 분석하였다. 종속변수는 '이직의사'인데 "이직의사 없음= 0, 이직의사 있음=1"로 코딩이 되었다.

케이스 처리 요약

가중되지 않은 케이스[a]		N	퍼센트
선택 케이스	분석에 포함	12657	62.5
	결측 케이스	7587	37.5
	합계	20244	100.0
비선택 케이스		0	.0
합계		20244	100.0

a. 가중값을 사용하는 경우에는 전체 케이스 수의 분류표를 참조하십시오.

종속변수 코딩

원래 값	내부 값
이직의사없음	0
이직의사있음	1

다음 범주형 변수 코딩 결과는 참조범주 및 기준값을 표시해준다. '직장소재지'의 경우 "서울=0, 경기강원=1"이 파라미터 코딩 (1)로 지정되었다. 이는 뒤의 출력결과의 '직장소재지(1)'을 의미한다. 또한 '직장소재지(2)'는 파라미터 코딩 (2)와 같아서 "서울=0, 영남=1", '직장소재지(3)'은 "서울=0, 호남=1", '직장소재지(4)'는 "서울=0, 충청=1"이 된다.

범주형 변수 코딩

		빈도	파러미터 코딩			
			(1)	(2)	(3)	(4)
문9.직장 소재지	서울	4197	.000	.000	.000	.000
	경기강원	2956	1.000	.000	.000	.000
	영남	3048	.000	1.000	.000	.000
	호남	1270	.000	.000	1.000	.000
	충청	1186	.000	.000	.000	1.000
문5.고용형태	상용직	11393	.000	.000		
	임시직	1188	1.000	.000		
	일용직	76	.000	1.000		
문22.노동조합 존재여부	노동조합이 없음	8947	.000			
	노동조합이 있음	3710	1.000			
문5-1.근로형태	정규직	11512	.000			
	비정규직	1145	1.000			
문2.일자리 형태	하루 종일 일하는 일반직장	12281	.000			
	시간단위로 일하는 파트타임이나 아르바이트	376	1.000			

블록 0: 시작 블록

분류표[a,b]

감시됨			예측		
			이직의사		
			이직의사없음	이직의사있음	분류정확 %
0 단계	이직의사	이직의사없음	9120	0	100.0
		이직의사있음	3537	0	.0
	전체 퍼센트				72.1

a. 모형에 상수항이 있습니다.

b. 절단값은 .500입니다.

시작 전 단계에서 모두 '이직의사없음'으로 분류해도 기본적인 분류정확도가 72.1%이므로 각 단계별로 변수가 추가됨에 따라 분류정확도가 개선되어야 할 것이다.

여기서 출력결과가 생략하지만 변수선택은 모두 6 단계까지 진행된다. 아래의 모형요약표에서 '−2Log우도' 는 카이제곱 분포를 따르는 통계량이며 Akaike 정보기준 함수식(=−2Log우도+$2k$)의 앞부분이다. 마지막 6 단계에서 AIC=12727.8(=12715.8+2×6)으로 전 단계 AIC=12739.4 (=12729.4+2×5)보다 줄어들었으나 그 차이가 매우 적어 다음 7단계에서는 AIC가 줄어들지 않았음을 추측할 수 있다.

모형 요약

단계	-2 Log 우도	Cox와 Snell의 R-제곱	Nagelkerke R-제곱
1	13417.795[a]	.117	.169
2	12935.186[a]	.150	.217
3	12799.625[a]	.159	.230
4	12747.728[a]	.163	.235
5	12729.428[a]	.164	.236
6	12715.837[a]	.165	.238

a. 모수 추정값이 .001보다 작게 변경되어 계산반복수 5에서 추정을 종료하였습니다.

= Hosmer와 Lemeshow 검정 =

단계	카이제곱	자유도	유의확률
1	26.883	2	.000
2	24.256	8	.002
3	18.669	8	.017
4	22.596	8	.004
5	30.609	8	.000
6	23.351	8	.003

각 단계별로 분류정확도와 포함된 변수들의 계수, 오즈비 등이 출력된다. 분류정확도는 1단계 전체 74.8%에서 최종 75.3%로 소폭 증가하였지만 이직의사 있는 사람을 예측하는 정확도는 10%p 증가하였다.

최종단계에 포함된 변수를 보면 '직장만족도', '급여', '직장소재지', '상용직여부', '노동조합 여부', '정규직 여부' 등이 이직의사에 영향을 주었다.

분류표[a]

감시됨			예측		
			이직의사		
			이직의사없음	이직의사있음	분류정확 %
1 단계	이직의사	이직의사없음	8778	342	96.3
		이직의사있음	2850	687	19.4
	전체 퍼센트				74.8
2 단계	이직의사	이직의사없음	8459	661	92.8
		이직의사있음	2538	999	28.2
	전체 퍼센트				74.7
3 단계	이직의사	이직의사없음	8483	637	93.0
		이직의사있음	2517	1020	28.8
	전체 퍼센트				75.1
4 단계	이직의사	이직의사없음	8488	632	93.1
		이직의사있음	2513	1024	29.0
	전체 퍼센트				75.2
5 단계	이직의사	이직의사없음	8489	631	93.1
		이직의사있음	2500	1037	29.3
	전체 퍼센트				75.3
6 단계	이직의사	이직의사없음	8490	630	93.1
		이직의사있음	2491	1046	29.6
	전체 퍼센트				75.3

a. 절단값은 .500입니다.

방정식에 포함된 변수

		B	S.E,	Wals	자유도	유의확률	Exp(B)
1 단계[a]	직장만족도	-1.045	.029	1311.754	1	.000	.352
	상수항	2.645	.099	717.687	1	.000	14.079
2 단계[b]	직장만족도	-1.014	.029	1199.372	1	.000	.363
	로그월급여	-1.174	.055	451.534	1	.000	.309
	상수항	8.468	.298	810.103	1	.000	4761.358
6 단계[f]	상용직여부			26.569	2	.000	
	상용직여부(1)	.403	.079	26.354	1	.000	1.497
	상용직여부(2)	.316	.268	1.388	1	.239	1.371
	정규직여부(1)	.303	.082	13.734	1	.000	1.354
	직장소재지			118.908	4	.000	
	직장소재지(1)	-.173	.058	8.933	1	.003	.841
	직장소재지(2)	-.549	.060	83.010	1	.000	.578
	직장소재지(3)	-.602	.083	52.748	1	.000	.548
	직장소재지(4)	-.008	.079	.010	1	.920	.992
	직장만족도	-1.015	.030	1177.426	1	.000	.362
	로그월급여	-1.051	.062	291.017	1	.000	.350
	노동조합유무(1)	-.234	.053	19.806	1	.000	.791
	상수항	8.057	.334	582.367	1	.000	3155.968

a. 변수가 1: 단계에 진입했습니다 직장만족도. 직장만족도.

b. 변수가 2: 단계에 진입했습니다 로그월급여. 로그월급여.

c. 변수가 3: 단계에 진입했습니다 직장소재지. 직장소재지.

d. 변수가 4: 단계에 진입했습니다 상용직여부. 상용직여부.

e. 변수가 5: 단계에 진입했습니다 노동조합유무. 노동조합유무.

f. 변수가 6: 단계에 진입했습니다 정규직여부. 정규직여부.

출력결과를 표로 요약하면 다음과 같다.

(이직의사 없음=0, 있음=1)

변수설명	자료변수명	B	유의확률	Exp(B)
상용직	상용직여부		.000	
상용(=0) 임시(=1)	상용직여부(1)	.403	.000	1.497
상용(=0) 일용(=1)	상용직여부(2)	.316	.239	1.371
정규직(=0) 비정규직(=1)	정규직여부(1)	.303	.000	1.354
직장소재지	직장소재지		.000	
서울(=0) 경기강원(=1)	직장소재지(1)	-.173	.003	.841
서울(=0) 영남(=1)	직장소재지(2)	-.549	.000	.578
서울(=0) 호남(=1)	직장소재지(3)	-.602	.000	.548
서울(=0) 충청(=1)	직장소재지(4)	-.008	.920	.992
직무만족도	직장만족도	-1.015	.000	.362
로그월급여	로그월급여	-1.051	.000	.350
노조(무=0, 유=1)	노동조합유무(1)	-.234	.000	.791
상수		8.057	.000	3155.968

결과를 해석하면 이직의사에 상용직여부가 영향을 유의하게 영향을 준다. 구체적으로 상용직에 비해 임시직이 이직 승산이 1.497배 증가하고(유의확률<0.001) 상용직에 비해 일용직은 1.37배 증가하나 통계적으로 유의하지 않다. 정규직에 비해 비정규직의 이직의사 오즈가 1.354배 높았으며, 서울에 비해 다른 지역에 있는 경우 이직의사 오즈가 낮다고 할 수 있다. 직무만족도가 높으면 이직의사 오즈가 낮아지며, 월급여도 유사한 결과를 보여준다. 노동조합이 있는 경우 이직의사 오즈가 낮아진다고 할 수 있다.

참고적으로 전일제 근무여부는 이직의사에 유의하게 영향을 주지 않았지만 전일제 변수 하나만 개별적으로 로지스틱 회귀분석하면 유의하게 영향을 준다는 점은 다중공선성의 개념으로 이해할 수 있을 것이다. 즉, 시간제나 전일제 근무는 상용직, 정규직 변수와 높은 연관을 갖기 때문이다. 그러므로 이 결과를 가지고 전일제 근무자나 시간제 근무자가 동일한 이직의사를 갖는다고 결론내리는 것은 현명하지 못하다. ■