Journal of the Society of Korea Industrial and Systems Engineering
Vol. 35, No. 3, pp.118－128, September 2012.

이원 이항 계수치 자료의 로지스틱 회귀 분석

안 해 일†

서경대학교 산업공학과

A Logistic Regression Analysis of Two-Way Binary Attribute Data

Haeil Ahn†

Seokyeong University

An attempt is given to the problem of analyzing the two-way binary attribute data using the logistic regression model in order to find a sound statistical methodology. It is demonstrated that the analysis of variance (ANOVA) may not be good enough, especially for the case that the proportion is very low or high. The logistic transformation of proportion data could be a help, but not sound in the statistical sense. Meanwhile, the adoption of generalized least squares (GLS) method entails much to estimate the variance-covariance matrix. On the other hand, the logistic regression methodology provides sound statistical means in estimating related confidence intervals and testing the significance of model parameters. Based on simulated data, the efficiencies of estimates are ensured with a view to demonstrate the usefulness of the methodology.

Keywords : Logistic, Binary Attribute Data, Logit Function, Odds Ratio, Dichotomous Response

1. 서 론

생산현장에서 공정의 성능을 분석하는 데 있어서는 조업조건에 따른 제조공정의 수율, 적합품의 비율, 부적합품의 비율 등 백분비 자료를 분석해야 하는 경우가 자주 발생되고 있다. 부적합률이 매우 낮아 0에 가깝거나 수율(yield rate)이 매우 높아 1에 근접하는 경우도 비교 분석해야 하는 경우가 많다. 예를 들어 국내의 어느 한 제조업체에서는 태블렛 PC 또는 스마트폰에 장착되는 디스플레이 장치 TSP(touch screen panel)를 생산하는데 매출이 연간 약 3,000~4,000억에 달하고 있다. 현재 수율이 약 93% 정도인데 앞으로 이 수율이 1%만 증가되어도 연간 수익이 수십억 원이 증대되는 효과를 거둘 수 있다고 한다.

수율의 증대를 위해서는 수율에 영향을 주는 영향 인자들의 수준 즉 조업조건에 대한 반응으로서의 수율 또는 부적합률의 변화를 검토 분석해야 한다. 조업조건에서의 반응 자료는 흔히 계수치 자료(attribute data, count data) 또는 이분법 반응자료(dichotomous response data)라 불리며 대학 교재 등 많은 문헌[2, 3]에 분산분석법(ANOVA)이 주로 사용되고 있다. 그런데 실제로 분산분석법을 적용해 보면 사리에 맞지 않는 경우가 자주 발생된다. 예를 들자면 최적 조업조건에서 수율의 점 추정치가 100%를 초과하여 초과달성이 가능하다는 해석을 해야 하는 경우가 있으며 구간추정에서는 부적합품의 비율의 하한이 음수가 되기도 한다. 이러한 비합리적인 추정결과로 인해 많은 분석자들로 하여금 분산분석 방법론을 신뢰하지 않게 하는 결과를 가져오게 된다.

이러한 문제는 오래 전부터 문헌에서 지적된 바 있는

논문접수일 : 2012년 07월 05일　　게재확정일 : 2012년 09월 03일
† 교신저자 hiahn@skuniv.ac.kr

것으로 보여 진다. 예를 들어 다구치[9]는 비율이 20%보다 낮거나 80%를 넘는 경우 가법성(additivity)이 성립되지 않는 다고 언급하고 오메(Ω)가 변환방법을 사용하였다. 하지만 이 또한 바람직한 해결책이라고 보기 어렵다.

이항 계수치 반응 자료를 수집한 후에 개선된 조업조건은 쉽게 찾을 수 있다고 생각한다. 하지만 실제로 적용하여 보면 최적 조업조건이라고 생각하였던 조업조건이 다음 실험에서는 다른 결과가 나오기도 하며 최악의 경우 정반대의 결과가 나오기도 한다. 그 이유는 현재의 조업조건과 비교하였을 때 어느 정도 자신할 수 있는지에 대한 신뢰구간 정보를 무시한 상태에서 평균 추정치만을 비교하는 경우가 많기 때문이다.

계수치 자료의 비교분석에 있어서는 변별력(discriminatory power)이 우수한 분석도구를 사용하여 검정과 신뢰구간 추정을 해보아야 모수 추정치의 신뢰수준과 자료의 추가 수집의 필요성 여부도 판단하기가 용이해 진다.

본 연구에서는 사리에 맞지 않는 계수치 자료 분석의 사례와 그 원인에 대하여 살펴보고 로지스틱 회귀(logistic regression)분석 방법론이 대안으로서 채택이 될 수 있는가를 고찰하고자 한다. 사례를 들어 자료를 분석하였을 때의 결과를 함께 제시하고 다원배치 계수치 자료의 분석 시 로지스틱 회귀 분석방법론의 적용법과 유용성에 대하여 논하고자 한다.

2. 백분비 자료의 분석

2.1 수율자료

백분비 자료로서 다음과 같은 이원 수율(yield rate)자료인데 여기서 A인자는 제조설비, B인자는 제조방법으로 이산적인 질적 변수이다.

<표 1> 이원 수율자료(%)

	A_1	A_2	A_3	A_4	평균
B_1	18.2	65.0	68.0	41.5	48.175
B_2	47.4	73.8	79.5	61.5	65.55
B_3	79.8	80.1	87.8	77.9	81.4
B_4	92.1	93.8	93.4	92.7	93.0
평균	59.375	78.175	82.175	68.4	72.031

전형적인 반복이 없는 2원배치의 문제로 간주하면 다음과 같은 자료구조 모형을 가정한다.

$$p_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \quad i = 1, 2, 3, 4 \quad j = 1, 2, 3, 4 \qquad (1)$$
$$\varepsilon_{ij} \sim \ni(0, \sigma_e^2), \quad \sum_{i=1}^{4} a_i = 0, \quad \sum_{j=1}^{4} b_j = 0$$

여기서 인자의 각 수준에서의 주 효과는 확정적이며 오차는 독립성과 등분산성을 갖는 것으로 가정한다. 교호작용을 고려할 수 없기 때문에 최적 수준조합은 최대치의 결과를 나타내는 A_2B_4가 아니라 A인자의 최대 수준과 B인자의 최대 수준에서 결정 된다. 즉, A_3B_4가 최적 수준이 되며 점 추정치는 다음과 같이 구해진다.

$$\hat{p}(A_3B_4) = \hat{p}(A_3) + \hat{p}(B_4) - \hat{p} \qquad (2)$$
$$= 82.175 + 93.0 - 72.031 = 103.144(\%)$$

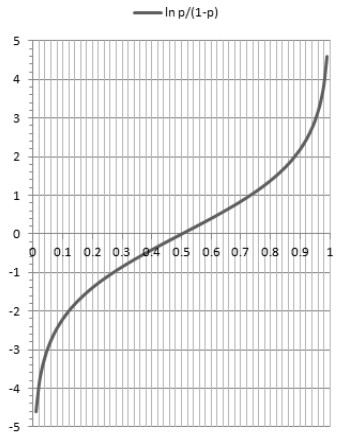
<표 2> 수율자료 Minitab 실행결과

```
새로운 관측치에 대한 예측치
새로운 관측치        적합지        SE
          1       103.144   7.07867
      95% CI                95% PI
(87.1307, 119.157)    (74.1176, 132.170)
새로운 관측치에 대한 예측 변수의 값
새로운 관측치  A  B
          1    3  4
```

문제는 점 추정치가 100%를 초과하는 것으로 나타난다. <표 2>와 같이 신뢰구간(CI)이나 예측구간(PI)의 상한도 물론 100%를 초과한다.

2.2 로지스틱 변환

로지스틱 변환(logistic transformation)이란 로짓(logit)이라고도 하며 <그림 1>과 같이 승산(odds)에 대수를 취한 값으로 계산한다.



<그림 1> 로짓(logit) 변환함수

다구치[9]에서는 여기에 10을 더 곱한 값을 사용하고 오메가(Ω) 변환이라고 하였다. 여기서는 로짓(logit) 변환만을 고려한다.

<표 3> 로짓 변환된 수율 자료

	A_1	A_2	A_3	A_4	평균
B_1	-1.503	0.6190	0.7538	-0.3433	-0.11838
B_2	-0.104	1.0356	1.3553	0.4684	0.688825
B_3	1.374	1.3926	1.9736	1.2600	1.50005
B_4	2.456	2.7166	2.6498	2.5415	2.5910
평균	0.55575	1.44095	1.6831	0.9817	1.1654

변환된 자료의 구조는 다음과 같이 가정한다.

$$r_{ij}=\ln[p_{ij}/(1-p_{ij})]=\mu+a_i+b_j+\varepsilon_{ij} \quad (3)$$
$$i=1,2,3,4 \quad j=1,2,3,4 \quad \varepsilon_{ij}\sim \ni(0,\sigma_e^2)$$
$$\sum_{i=1}^{4}a_i=0, \quad \sum_{j=1}^{4}b_j=0$$

교호작용이 없는 경우이므로 변환치가 최대가 되는 각 인자의 최적 수준조합은 역시 A_2B_4가 아니라 A_3B_4가 되며 추정치는 다음과 같이 구해진다.

$$\tilde{r}(A_3B_4)=\tilde{r}(A_3)+\tilde{r}(B_4)-\tilde{r} \quad (4)$$
$$=1.6831+2.5910-1.1654=3.1087$$

이 수치를 역변환하면

$$\tilde{p}=\frac{e^{3.1087}}{1+e^{3.1087}}\equiv\frac{1}{1+e^{-3.1087}}=0.95725=95.725(\%) \quad (5)$$

와 같이 계산된다. <표 4>의 결과를 이용하면 신뢰구간과 예측구간도 계산이 가능하다.

신뢰구간(CI) (0.918909, 0.977902)
예측구간(PI) (0.866930, 0.987175)

<표 4> 변환자료 Minitab 실행결과

새로운 관측치에 대한 예측치		
새로운 관측치	적합지	SE
1	3.10877	0.30110
95% CI		95% PI
(2.42762, 3.78992)		(1.87408, 4.34346)

구간 추정치의 상한이 1을 초과하지 않는다. 이와 같이 좀 더 수긍할 수 있는 결과를 얻게 된다. 문제는 시행횟수와 발생횟수에 대한 정보가 없으며 그러한 정보가 있더라도 이러한 모형을 적용할 때에는 무시될 수밖에 없다. 만약 시행횟수에 대한 정보가 있으면 반복횟수로 간주되어 교호작용을 구할 수도 있다. 따라서 이론적으로 좀 더 바람직한 분석방법론이 요구된다.

3. 계수치 자료의 분석

3.1 부적합률 자료

두 가지 질적 인자 기계(A)와 열처리방법(B)에 대한 계수치 자료가 <표 5>와 같이 주어져 있다. 여기서 수준수는 A인자 $l=4$, B인자 $m=2$, 반복수 $r=120$, 총 실험횟수 $N=960$ 인 반복 있는 이원 배치로 간주할 수 있다.

<표 5> 이항 계수치 자료

기계 열처리	A_1		A_2		A_3		A_4		계
	적	부	적	부	적	부	적	부	
B_1	115	5	108	12	117	3	100	20	$T_{1.}=40$
B_2	110	10	100	20	112	8	98	22	$T_{2.}=60$
계	$T_{1..}=15$		$T_{2..}=32$		$T_{3..}=11$		$T_{4..}=42$		$T=100$

현재의 조업조건은 A_1B_1로서 대조군(control group)이며 여타의 조업조건인 실험군(experiment group)들에 대해 유의성을 검정하고 최적 조업조건에서 부적합률을 추정하여 개선된 조업조건을 찾는 문제이다.

3.2 분산분석

처리 수준조합은 완전 확률화에 의해 결정되었으나 수준조합 내에서의 관측치는 확률화가 되어 있지 않으므로 자료의 구조는 다음과 같이 상정하는 것이 일반적이다.

$$y_{ijk}=\mu+a_i+b_j+\varepsilon_{(1)ij}+\varepsilon_{(2)ijk} \quad (6)$$
$$y_{ijk}=\begin{cases}0: \text{적합}\\ 1: \text{부적합}\end{cases}$$

$\varepsilon_{(1)ij}$는 1차 단위의 오차로서 교호작용효과($A\times B$)와 교락되어 있는 분산성분이고 $\varepsilon_{(2)ij}$는 오차이다.

3.2.1 분산분석표

반복이 있는 이원 배치와 같이 계산될 수 있으며 특히 y_{ijk}는 0 또는 1를 나타내는 변수이고 $y_{ijk}^2=y_{ijk}$이므로 분산분석방업을 적용 계산할 수 있다. 정리된 분산분석표는 다음 <표 6>과 같다.

<표 6> 분산분석표

출처	SS	ν	MS	F_0	$p-value$
A	2.641	3	0.8803	34.79^{**}	<0.01
B	0.416	1	0.4160	16.44^{*}	<0.05
E_1	0.076	3	0.0253	0.28	
E_2	86.450	952	0.0908		
T	89.583	959			

우선 1차단위의 오차 $E_1(A\times B)$의 유의성 검정에서 $F_0=0.28<1$이므로 유의하지 않은 것으로 판단하여 풀링하기로 결정하였다면 분산분석표는 <표 7>과 같이 다시 작성된다.

<표 7> 풀링된 분산분석표

출처	SS	ν	MS	F_0	$p-value$
A	2.641	3	0.8803	9.716**	<0.01
B	0.416	1	0.4160	4.592*	<0.05
E	86.526	955	0.906		
T	89.583	959			

분석결과 A, B인자 모두 유의하다고 판단된다.

3.2.2 부적합률 추정

개선된 조업조건은 유일하게 A_3B_1이다. 교호작용이 무시되는 경우이므로 비율의 추정은

$$\tilde{p}(A_3B_1)=\tilde{p}(A_3)+\tilde{p}(B_1)-\tilde{p}$$
$$=4.58+8.33-\left(\frac{100}{960}\times 100\right)=2.49(\%)$$

유효반복수는 다음과 같이 계산된다.

$$n_e=\frac{lmr}{l+m-1}=\frac{4\times 2\times 120}{4+2-1}=192$$

95% 신뢰구간 폭은 다음과 같다.

$$t_{\alpha/2}(\nu_E)\sqrt{\frac{MS_E}{n_e}}=t_{0.025}(955)\sqrt{\frac{0.0906}{192}}$$
$$=1.96\times 0.0217=0.0425$$

따라서 $p(A_3B_1)$에 대한 95% 신뢰구간은

$$p(A_3B_1):2.49\pm 4.25(\%)=(-1.76\%,\ 6.74\%)$$

신뢰구간의 하한이 음수로 계산된다. 통계적으로는 신뢰구간이 0을 포함하므로 부적합률을 0으로 보아도 된다는 귀무가설이 받아들여 질 수 있다. 통계적인 유의성을 판별하기가 매우 어려운 상황이다.

이러한 결과가 나오는 주된 이유는 (1) 우선 발생 도수가 너무 작고(<5). (2) 수준조합에서의 사건 발생비율이 0에 가깝거나 1에 매우 가까울 경우에는 시행횟수가 많아야만 하고 정규근사 조건 $n_{ij}p_{ij}\geq 5$와 $n_{ij}(1-p_{ij})\geq 5$을 만족해야만 정규분포 근사가 가능한데 이 경우 그렇지 못하며 (3) 비율이 0.5가 아닐 때 이항분포는 비대칭적인 분포를 하는데 대칭적인 정규근사를 하였기 때문인 것으로 생각된다. 또한 (4) 분산분석 모형에서는 각 수준조합에서의 오차변동은 등분산을 갖는 대칭적인 정규분포를 가정하고 있는데 실제로 표본분포의 경우 등분산성 조건이 만족되지 않는다.

엄밀한 의미에서 계수치 자료의 분석에 분산분석법이 사용될 수 없을 것으로 생각된다.

3.3 로짓 변환분석

로짓 변환을 통해 자료를 $r_{ij}=\ln[p_{ij}/(1-p_{ij})]$로 계산한 결과는 <표 8>과 같다.

<표 8> 로짓(오메가) 변환치

열처리기계	B_1	B_2	계	평균
A_1	-3.1355	-2.3979	-5.5334	-2.7667
A_2	-2.1972	-1.6094	-3.8066	-1.9033
A_3	-3.6636	-2.6391	-6.3027	-3.15135
A_4	-1.6094	-1.4939	-3.1033	-1.55265
계	-10.6057	-8.1403	-18.746	
평균	-2.65143	-2.03508		-2.34325

이 자료구조는 반복 없는 2원 배치 자료로 간주할 수 있다. 다음과 같은 모형을 적용한다.

$$r_{ij}=\ln[p_{ij}/(1-p_{ij})]=\mu+a_i+b_j+\varepsilon_{ij} \quad (7)$$
$$i=1,2,3 \quad j=1,2 \quad \varepsilon_{ij}\sim \ni(0,\sigma_e^2)$$
$$\sum_{i=1}^{4}a_i=0,\quad \sum_{j=1}^{2}b_j=0$$

<표 9> 로짓 변환치에 대한 이원 분산분석

출처	ν	SS	MS	F_0	$p-value$
A	3	2.9979	0.99929	4.82*	0.029
B	3	16.0775	5.35917	25.86**	0.000
E	9	1.8651	0.20723		
T	15	20.9405			

분석결과 A, B인자 모두 유의하다고 판단할 수 있다. 부적합률은 망소특성 자료이며 교호작용을 구할 수 없는 경우로서 A인자의 최저수준 B인자의 최저수준에서 최적이므로 최적 수준조합은 A_3B_1으로 평균에 대한 구간추정은

$$\tilde{r}(A_3B_1)=\tilde{r}(A_3)+\tilde{r}(B_1)-\tilde{r}$$
$$=-3.15135-2.65142+2.34325=-3.45952$$
$$n_e=\frac{lm}{l+m-1}=\frac{3\times 3}{4+2+1}=1.2857$$

$r(A_3B_1)$에 대한 95% 신뢰구간은 다음과 같다.

$$\begin{aligned}&-3.45952 \pm t_{0.025(9)}\sqrt{MS_E/1.2857}\\&= -3.45952 \pm 2.262 \times 0.40147\\&= (-4.36749, -2.55139)\end{aligned}$$

변환치에 대한 최적 수준조합 A_3B_1에서의 점 추정과 구간 추정은 다음과 같다.

$$\widetilde{p_{31}} = \frac{1}{1+e^{-r_{31}}} = \frac{1}{1+e^{3.45952}} = 0.03049$$

$p(A_3B_1)$에 대한 95% CI: (0.01252, 0.07233)

하한에 음의 수가 나타나지 않는다는 점에서 자료변환의 1차적인 목표는 달성되었지만 하지만 자료의 시행횟수에 대한 정보를 무시하고 있기 때문에 관측치가 가지고 있는 정보 중 많은 부분의 손실이 발생된다.

3.4 이항 로지스틱 회귀

이항 로지스틱 회귀 이론은 많은 문헌에 존재하며 Kleinbaum and Klein[6]과 Montgomery[8] 등에 비교적 상세히 소개되어 있다. 로지스틱 회귀분석 방법론은 그간 이분법적(dichotomous) 반응 자료[5, 10]에 대한 분할표(contingency table) 검정이나 다분법적(polychotomous) 반응으로서의 범주형(categorical) 자료[5, 11]의 분석에 사용되고 있다고 생각된다.

3.4.1 자료 구조 모형

일반적으로 분산분석 모형을 회귀분석 모형으로 변환하는 방법은 다음과 같다. 편의상 반복이 있는 이원배치 모수모형을 다음과 같이 정의한다.

$$\begin{aligned}&y_{ijk} = \alpha + \beta_i + \gamma_j + \delta_{ij} + \varepsilon_{ijk} \quad \varepsilon_{ijk} \sim N(0, \sigma^2) \\&i = 1, 2, \cdots, l \quad j = 1, 2, \cdots, m \quad k = 1, 2, \cdots, r\end{aligned} \tag{8}$$

제약식 $\sum_{i=1}^{l}\beta_i = \sum_{j=1}^{m}\gamma_j = \sum_{i=1}^{l}\delta_{ij} = \sum_{j=1}^{m}\delta_{ij} = 0$이 필요하다. 이에 대응하는 회귀모형은

$$\begin{aligned}y_{ijk} &= \alpha + \sum_{i=1}^{l}\beta_i x_i + \sum_{j=1}^{m}\gamma_j x_{l+j} \\&\quad + \sum_{i=1}^{l}\sum_{j=1}^{m}\delta_{ij}x_i x_{l+j} + \varepsilon_{ijk}\end{aligned} \tag{9}$$

$$x_u = \begin{cases}1, & u = i\\0, & u \neq i, u = 1, 2, \cdots, l\end{cases}$$

$$x_{l+v} = \begin{cases}1, & v = j\\0, & v \neq j, v = 1, 2, \cdots, m\end{cases}$$

로 정의되며 여기에 제약식이 추가되어야 한다.

3.4.2 로지스틱 회귀모형

<표 5>의 예제에 대한 자료구조는 다음과 같다.

$$\begin{aligned}&g(p_{ij}) = \ln\left[p_{ij}/(1-p_{ij})\right] = \alpha + \beta_i + \gamma_j + \delta_{ij} \\&i = 1, 2, 3, 4 \qquad j = 1, 2\end{aligned} \tag{10}$$

여기에 일반적으로 다음 제약식을 추가한다. 즉,

$$\begin{aligned}&\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0,\ \gamma_1 + \gamma_2 = 0\\&\delta_{11} + \delta_{12} = 0,\ \delta_{21} + \delta_{22} = 0\\&\delta_{31} + \delta_{32} = 0,\ \delta_{41} + \delta_{42} = 0\\&\delta_{11} + \delta_{21} + \delta_{31} + \delta_{41} = 0\\&\delta_{12} + \delta_{22} + \delta_{32} + \delta_{42} = 0\end{aligned}$$

와 같은 식을 추가한다. 이 제약식 들을 모형에 직접 반영하는 몇 가지 방법 중 한 가지는 각 제약식의 첫 번째 계수를 0으로 설정하고 즉,

$$\beta_1 = \gamma_1 = 0, \quad \delta_{11} = \delta_{12} = 0, \quad \delta_{21} = \delta_{31} = \delta_{41} = 0$$

으로 간주하고 그에 따라 α의 값이 결정되도록 하면 α는 대조군의 평균을 의미하는 것이 된다.

로지스틱 회귀모형에서는 로짓(logit) 연결함수(link function) $g(.)$을 도입하여 오메가 변환과 유사하게 다음과 같이 정의한다.

$$\begin{aligned}g(p_{ij}) &= \ln\left[p_{ij}/(1-p_{ij})\right] = \mathbf{x}_{ij}'\boldsymbol{\beta} \\&= \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \gamma_2 x_6 \\&\quad + \delta_{22}x_2x_6 + \delta_{32}x_3x_6 + \delta_{42}x_4x_6 \\&i = 1, 2, 3, 4, \quad j = 1, 2\end{aligned} \tag{11}$$

이러한 모형은 일반화 선형모형(GLM)에 해당하며 설명변수는 가변수(dummy variable)이다.

$$\boldsymbol{X} = \begin{bmatrix}1&0&0&0&0&0&0&0\\1&0&0&0&1&0&0&0\\1&1&0&0&0&0&0&0\\1&1&0&0&1&1&0&0\\1&0&1&0&0&0&0&0\\1&0&1&0&1&0&1&0\\1&0&0&1&0&0&0&0\\1&0&0&1&1&0&0&1\end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix}\alpha\\\beta_2\\\beta_3\\\beta_4\\\gamma_2\\\delta_{22}\\\delta_{32}\\\delta_{42}\end{bmatrix} \tag{12}$$

설계행렬(design matrix) $\boldsymbol{X}$와 모수 벡터 $\boldsymbol{\beta}$는 식 (12)와 같다. $\boldsymbol{X}$행렬의 특정한 행을 나타내는 $\mathbf{x}_{ij}'$는 수준조합 A_iB_j에 해당하는 행벡터이다.

3.4.3 모형의 적합

로지스틱 회귀에서는 모형의 적합과정에서 최우 추정법을 직접 사용하고 또한 반복적인 알고리즘을 사용한다. 로지스틱 회귀의 추론 과정에서는 다음과 같은 이항

확률변수 Z를 고려한다.

$$Z=\begin{cases}1, & \text{출현}\\ 0, & \text{미출현}\end{cases} \tag{13}$$

확률은 $\Pr(Z=1)=p$, $\Pr(Z=0)=1-p$로 각각 정의한다. 2차원 배열로 주어진 $l\times m$개의 수준조합이 있다면 $\Pr(Z_{ij}=1)=p_{ij}$라고 할 때 결합 확률함수는 다음과 같다.

$$\begin{aligned}&\prod_{i=1}^{l}\prod_{j=1}^{m}p_{ij}^{z_{ij}}(1-p_{ij})^{1-z_{ij}} \\ &=\exp\left[\sum_{i}^{l}\sum_{j}^{m}z_{ij}\ln\left(\frac{p_{ij}}{1-p_{ij}}\right)+\sum_{i=1}^{l}\sum_{j=1}^{m}\ln(1-p_{ij})\right]\end{aligned} \tag{14}$$

$Y=\sum_{k=1}^{n_{ij}}Z_k$라 정의하면 Y는 이항분포이므로

$$\Pr(Y=y_{ij})=\binom{n_{ij}}{y_{ij}}p^{y_{ij}}(1-p)^{n_{ij}-y_{ij}} \tag{15}$$
$$y_{ij}=0,\ 1,\ 2,\ 3,\ \cdots,\ n_{ij}$$

$Y_{ij}\sim BIN(n_{ij},\ p_{ij})$ 로 정의하면 로그우도함수는

$$\begin{aligned}&l(p_{11},\cdots,p_{lm};\ y_{11},\cdots,y_{lm}) \\ &=\sum_{i}^{l}\sum_{j}^{m}\left[y_{ij}\ln\left(\frac{p_{ij}}{1-p_{ij}}\right)+n_{ij}\ln(1-p_{ij})+\ln\binom{n_{ij}}{y_{ij}}\right]\end{aligned} \tag{16}$$

이다. 로지스틱 회귀에서의 연결함수(link function) $g(\cdot)$를 $g(p_{ij})=\ln[p_{ij}/(1-p_{ij})]=\mathbf{x}_{ij}'\boldsymbol{\beta}$와 같이 표시하면 $\ln(1-p_{ij})=-\ln[1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})]$이므로 로그 우도함수는 다음과 같이 된다.

$$\begin{aligned}&l(p_{11},\cdots,p_{lm};y_{11},\cdots,y_{lm}) \\ &=\sum_{i}^{l}\sum_{j}^{m}\left[y_{ij}(\mathbf{x}_{ij}'\boldsymbol{\beta})-n_{ij}\ln(1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}))+\ln\binom{n_{ij}}{y_{ij}}\right]\end{aligned} \tag{17}$$

모수($\boldsymbol{\beta}$)에 대해 편미분으로 스코어 벡터(score vector)인 $\mathbf{u}$를 구하고 헤시안(Hessian)으로 정보행렬(information matrix) $\boldsymbol{I}(\boldsymbol{\beta})$를 구한다.

$\boldsymbol{\beta}$에 대한 최우 추정치 벡터 $\hat{\boldsymbol{\beta}}$은 다음과 같이 반복적으로 구할 수 있다.

$$\hat{\boldsymbol{\beta}}^{(k)}=\hat{\boldsymbol{\beta}}^{(k-1)}-[\mathbf{I}\ (\boldsymbol{\beta})^{(k-1)}]^{-1}\mathbf{u}^{(k-1)},\ k=1,2,3,\cdots \tag{18}$$

이것은 Newton-Raphson과 유사한 수치해석적인 탐색방법으로 일반화 추정식(generalized estimating equation; GEE)으로도 부른다.

여기서 정보행렬은 $\boldsymbol{I}=\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$에 해당하는 것으로서 $\hat{\boldsymbol{\beta}}$를 이 알고리즘의 최종 추정치라고 하면

$$E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta},\quad Var(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\sigma^2 \tag{19}$$

$\boldsymbol{V}$는 계수들 간의 분산-공분산 행렬이며 i번째 대각원소 V_{ii}는

$$V_{ii}=n_i\hat{p_i}(1-\hat{p_i}) \tag{20}$$

로 분산의 추정치가 된다. GLS의 결과와 일치한다. $\hat{\boldsymbol{\beta}}$를 구한 후 적합된 추정치는

$$\hat{p_{ij}}=\frac{\exp(\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}})}=\frac{1}{1+\exp(-\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}})} \tag{21}$$

각 수준조합에서 발생확률의 구간추정을 위해서는 수준조합을 예측변수들의 벡터로 표시하여야 한다. $\mathbf{x}'_0=[1,\ x_2^0,\ x_3^0,\ x_4^0,\ x_6^0]$을 회귀변수의 특정 값에서의 벡터라고 하면 $\mathbf{x}_0$에서의 추정치 p_0는 $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$로 추정될 수 있다. 또한

$$\begin{aligned}&E(\mathbf{x}_0'\hat{\boldsymbol{\beta}})=\mathbf{x}_0'\boldsymbol{\beta} \\ &Var(\mathbf{x}_0'\hat{\boldsymbol{\beta}})=\mathbf{x}_0'\,Var(\hat{\boldsymbol{\beta}})\mathbf{x}_0=\mathbf{x}_0'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\mathbf{x}_0\sigma^2\end{aligned} \tag{22}$$

이므로 $\mathbf{x}_0$에서의 발생확률 p_0의 신뢰구간의 하한과 상한은 다음과 같이 계산된다.

$$\begin{aligned}L(\mathbf{x}_0)&=\mathbf{x}_0'\hat{\boldsymbol{\beta}}-z_{\alpha/2}\sqrt{[\mathbf{x}_0'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\mathbf{x}_0]MS_E} \\ U(\mathbf{x}_0)&=\mathbf{x}_0'\hat{\boldsymbol{\beta}}+z_{\alpha/2}\sqrt{[\mathbf{x}_0'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\mathbf{x}_0]MS_E}\end{aligned} \tag{23}$$

여기서의 하한과 상한은 역 변환하면 발생률 p_0의 하한과 상한으로 계산될 수 있다. 즉,

$$\frac{\exp[L(\mathbf{x}_0)]}{1+\exp[L(\mathbf{x}_0)]}\le p_0\le\frac{\exp[U(\mathbf{x}_0)]}{1+\exp[U(\mathbf{x}_0)]} \tag{24}$$

이 과정에서 시행횟수가 반영된 추정결과를 얻게 되기 때문에 시행횟수가 많을수록 정밀도가 높은 즉, 신뢰구간이 더 좁은 추정치를 얻을 수 있다.

<표 10> 로지스틱 회귀 분석 표

로지스틱 회귀 분석 표

예측변수	계수	SE 계수	Z	P	승산비	95% CI 하한	95% CI 상한
상수	-3.13549	0.456832	-6.86	0.000			
x2 1	0.938270	0.548897	1.71	0.087	2.56	0.87	7.49
x3 1	-0.528067	0.742008	-0.71	0.477	0.59	0.14	2.53
x4 1	1.52606	0.518359	2.94	0.003	4.60	1.67	12.71
x6 1	0.737599	0.563726	1.31	0.191	2.09	0.69	6.31
x2x6 1	-0.149812	0.685842	-0.22	0.827	0.86	0.22	3.30
x3x6 1	0.286905	0.890840	0.32	0.747	1.33	0.23	7.64
x4x6 1	-0.622086	0.658366	-0.94	0.345	0.54	0.15	1.95

자료의 처리 시 SAS PROC LOGISTIC[4]나 Minitab [7]과 같은 패키지를 사용할 수 있다. <표 10>은 미니탭을 사용하여 식 (11)에 주어진 모형의 모수를 추정한 결과이다. 여기서 유의하지 않은 모수에 대한 정보도 얻을 수 있다.

3.4.4 유의하지 않은 모수 제거

교호작용에 해당하는 계수들 δ_{22}, δ_{32}, δ_{42}과 β_3가 유의하지 않은 것으로 판단하여 모형에서 제거하기로 결정하였다고 가정하면 적합된 모형은

$$g(\widehat{p_{ij}}) \equiv \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} = \hat{\alpha} + \hat{\beta}_2 x_2 + \hat{\beta}_4 x_4 + \hat{\gamma}_2 x_6 \qquad (25)$$
$$i = 1, 2, 3, 4 \quad j = 1, 2$$
$$\hat{\alpha} = -3.11675,\ \hat{\beta}_2 = 0.992392$$
$$\hat{\beta}_4 = 1.31582,\ \hat{\gamma}_2 = 0.465494$$

여기서 특기할 것은 현재의 조업조건(A_1B_1)에서

$$g(\widehat{p_{11}}) = \ln\frac{\widehat{p_{11}}}{1-\widehat{p_{11}}} = \hat{\alpha} = -3.11675$$
$$\Rightarrow \widehat{p_{11}} = \frac{1}{1+e^{-\hat{\alpha}}} = \frac{1}{1+e^{3.11675}} = 0.042422$$

개선된 조업조건(A_3B_1)에서

$$g(\widehat{p_{31}}) = \ln\frac{\widehat{p_{31}}}{1-\widehat{p_{31}}} = \hat{\alpha} = -3.11675$$
$$\Rightarrow \widehat{p_{31}} = \frac{1}{1+e^{-\hat{\alpha}}} = \frac{1}{1+e^{3.11675}} = 0.042422$$

와 같이 A_1 수준과 A_3 수준에서의 추정치가 동일하게 나타나고 있다. 이는 계수 β_1과 β_3를 0으로 설정하였기 때문인 것으로 생각된다.

<표 11> 수준조합에서의 발생비율 추정

수준		1	x_2	x_4	x_6	발생률	$\widehat{p_{ij}}$
A	B	α	β_2	β_4	γ_2	y_{ij}/n_{ij}	
1	1	1	0	0	0	5/120	0.042422
1	2	1	0	0	1	10/120	0.065912
2	1	1	1	0	0	12/120	0.106752
2	2	1	1	0	1	20/120	0.159915
3	1	1	0	0	0	3/120	0.042422
3	2	1	0	0	1	8/120	0.065912
4	1	1	0	1	0	20/120	0.141738
4	2	1	0	1	1	22/120	0.208262

주) 현재 조업조건 A_1B_1, 개선된 조업조건 A_3B_1.

3.4.5 유의하지 않은 모수 일부 포함

추정결과 계수 β_3가 유의하지 않은 것으로 나타나고는 있지만 적어도 하나의 β_i, $i = 2, 3, 4$ 계수가 유의하므로 인자 A가 유의하며 앞의 분산분석 결과와 비교를 위하여 모형에 β_3를 포함시켰다.

$$g(p_{ij}) = \ln[p_{ij}/(1-p_{ij})] = \mathbf{x}_{ij}'\boldsymbol{\beta} \qquad (26)$$
$$= \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \gamma_2 x_6$$
$$i = 1, 2, 3, 4, \quad j = 1, 2$$

이것은 교호작용만을 제외시킨 가법모형에 해당한다. 로지스틱 회귀로 추정한 결과가 <표 12>와 같다. 여러 번 반복실행 결과 로그우도가 최대가 되는 해를 구하였다는 사실을 확인할 수 있다.

<표 12> 가법모형 추정 결과

단계 로그 우도
0 -320.777
1 -305.145
2 -303.994
3 -303.986
4 -303.986
5 -303.986

연결 함수 : 로짓

반응 정보

변수	값	카운트
y	사건	100
	비사건	860
n	총계	960

로지스틱 회귀 분석 표

예측 변수	계수	SE 계수	Z	P	승산비	95% CI 하한	95% CI 상한
상수	-2.96446	0.297568	-9.96	0.000			
x2							
1	0.840005	0.328059	2.56	0.010	2.32	1.22	4.41
x3							
1	-0.328655	0.408445	-0.80	0.421	0.72	0.32	1.60
x4							
1	1.16344	0.316930	3.67	0.000	3.20	1.72	5.96
x6							
1	0.465669	0.218431	2.13	0.033	1.59	1.04	2.44

로그 우도 = -303.986

모든 기울기가 0인지 검정 : G = 33.582, DF = 4, P-값 = 0.000

적합도 검정

방법	카이-제곱	DF	P
Pearson	2.06985	3	0.558
이탈도	2.09367	3	0.553
Hosmer-Lemeshow	1.12680	2	0.569

카이자승 검정이나 이탈도의 p-value가 유의수준 $\alpha = 0.05$에 비해 매우 큰 것으로 보아 가정된 모형이 적합한 것으로 판단된다. 추정된 모형은 다음과 같이 정리된다.

$$g(\hat{p_{ij}}) = \ln\left[\hat{p_{ij}}/(1-\hat{p_{ij}})\right] = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} \quad (27)$$
$$= \hat{\alpha} + \hat{\beta_2}x_2 + \hat{\beta_3}x_3 + \hat{\beta_4}x_4 + \hat{\gamma_2}x_6$$
$$\hat{\alpha} = -2.96446,\ \hat{\beta_2} = 0.840005,\ \hat{\beta_3} = -0.328655$$
$$\hat{\beta_4} = 1.16344, \hat{\gamma_2} = 0.465669$$

발생확률은 역변환을 하여 추정이 가능하다. 즉,

$$\hat{p_{ij}} = \frac{\exp(\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}})} = \frac{1}{1+\exp(-\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}})} \quad (28)$$

와 같은 방식으로 계산이 가능하다. 그런데 다음 조건이 만족되어야 하므로

$$\beta_1 = \gamma_1 = 0, \quad \delta_{11} = \delta_{12} = 0 \quad (29)$$
$$\delta_{21} = \delta_{31} = \delta_{41} = 0$$

현재의 조업조건(A_1B_1)에서의 추정치는

$$g(\hat{p_{11}}) = \ln\frac{\hat{p_{11}}}{1-\hat{p_{11}}} = \hat{\alpha} = -2.96446$$
$$\Rightarrow \hat{p_{11}} = \frac{1}{1+e^{-\hat{\alpha}}} = \frac{1}{1+e^{2.96446}} = 0.049057$$

개선된 조업조건(A_3B_1)에서의 추정치는

$$g(\hat{p_{31}}) = \ln\frac{\hat{p_{31}}}{1-\hat{p_{31}}} = \hat{\alpha} + \hat{\beta_3} = -3.29312$$
$$\Rightarrow \hat{p_{31}} = \frac{1}{1+e^{-(\hat{\alpha}+\hat{\beta_3})}} = \frac{1}{1+e^{3.29312}} = 0.035808$$

한 가지 예를 더 든다면 A_3B_2에서는

$$g(\hat{p_{32}}) = \ln\frac{\hat{p_{32}}}{1-\hat{p_{32}}} = \hat{\alpha} + \hat{\beta_3} + \hat{\gamma_2} = -2.82745$$
$$\Rightarrow \hat{p_{32}} = \frac{1}{1+e^{-(\hat{\alpha}+\hat{\beta_3}+\hat{\gamma_2})}} = \frac{1}{1+e^{2.82745}} = 0.055859$$

모든 수준조합에서의 발생확률은 <표 13>과 같다.

<표 13> 수준조합에서의 발생비율 추정

수준		1	x_2	x_3	x_4	x_6	발생률	$\hat{p_{ij}}$
A	B	α	β_2	β_3	β_4	γ_2	y_{ij}/n_{ij}	
1	1	1	0	0	0	0	5/120	0.049057
1	2	1	0	0	0	1	10/120	0.075943
2	1	1	1	0	0	0	12/120	0.106742
2	2	1	1	0	0	1	20/120	0.159924
3	1	1	0	1	0	0	3/120	0.035808
3	2	1	0	1	0	1	8/120	0.055859
4	1	1	0	0	1	0	20/120	0.141726
4	2	1	0	0	1	1	22/120	0.208274

주) 현재 조업조건 A_1B_1, 개선된 조업조건 A_3B_1.

다음 <표 14>는 미니탭을 이용한 발생률의 추정치와 95% 신뢰구간이다.

<표 14> 발생률 추정치와 신뢰구간

수준		발생률	추정치	95% 신뢰구간	
A	B	y_{ij}/n_{ij}	$\hat{p_{ij}}$	하한	상한
1	1	5/120	0.049057	0.027985	0.084614
1	2	10/120	0.075943	0.045240	0.124758
2	1	12/120	0.106742	0.070832	0.157767
2	2	20/120	0.159924	0.111823	0.223510
3	1	3/120	0.035808	0.018859	0.066951
3	2	8/120	0.055859	0.030580	0.099882
4	1	20/120	0.141726	0.098299	0.200080
4	2	22/120	0.208274	0.152410	0.277901

주) 현재 조업조건 A_1B_1, 개선된 조업조건 A_3B_1.

이 경우 발생비율이 낮음에도 불구하고 신뢰구간이 비대칭적이며 하한이 음수로 추정되는 현상이 없을 뿐만 아니라 오차 분산의 동일성이나 독립성의 가정이 없이도 적합(fitting)된 추정치를 구할 수 있다.

발생비율 추정치는 0.035808로 실현 비율 3/120 = 0.025 보다 다소 높게 추정되고 있다. 이것은 로지스틱곡선으로 적합된(fitting)결과로서 통계적인 관점에서는 더 바람직한 결과이다.

3.4.6 조업조건의 유의성 검정

현재의 조업조건(A_1B_1) 대비 개선된 조업조건(A_3B_1)의 OR(odds ratio)는 다음과 같이 계산된다.

$$OR = \frac{\widetilde{p_{31}}/(1-\widetilde{p_{31}})}{\widetilde{p_{11}}/(1-\widetilde{p_{11}})} = \frac{e^{\hat{\alpha}+\hat{\beta_3}}}{e^{\hat{\alpha}}} = e^{\hat{\beta_3}} = 0.72$$

이것은 부적합률이 현재의 72% 수준으로 저감될 수 있다는 의미로 해석되지만 이 승산비의 신뢰구간은 (0.32, 1.60)으로 나타나고 있기 때문에 즉, 신뢰구간이 1을 포함하고 있기 때문에 통계적으로 유의하다고 할 수는 없다.

이와 같이 현재의 조업조건에서의 결과를 기준으로 다른 조업조건 들에서의 OR(odds ratio)을 구할 수 있으며 OR의 점 추정치와 구간 추정치로 비율의 차이의 추정과 신뢰구간으로 별도의 검정과정 없이 손쉽게 유의성을 검정할 수 있다.

3.5 모의실험

앞에서 유의하지 않다는 결과가 나온 주된 이유는 시행횟수가 적기 때문이라고 생각된다. 따라서 시행횟수가 많을 때의 효과를 확인하기 위하여 발생비율은 동일하지

만 시행횟수와 발생횟수가 4배로 큰 경우의 일종의 모의자료를 만들었다.

<표 15> 이항 계수치 모의자료

기계 열처리		B_1	B_2	계
A_1	부적합	20	40	60
	적합	460	440	900
	소계	480	480	960
A_2	부적합	48	80	128
	적합	432	400	832
	소계	480	480	960
A_3	부적합	12	32	44
	적합	468	448	916
	소계	480	480	960
A_4	부적합	80	88	168
	적합	400	392	792
	소계	480	480	960
계	부적합	160	240	400
	적합	1760	1680	3440
	소계	1920	1920	3840

3.5.1 분산분석법의 적용

자료구조모형은 모형은 식 (6)에서와 동일하게 가정하였다. 반복이 있는 이원 배치와 같이 계산될 수 있으며 특히 y_{ijk}는 0 또는 1를 나타내는 변수이고 $y_{ijk}^2 = y_{ijk}$이므로 다음과 같이 계산할 수 있다.

$$CT = T^2/lmr = 400^2/(4\times 2\times 480) = 41.667$$

$$S_T = \sum_{i=1}^{4}\sum_{j=1}^{2}\sum_{k=1}^{480} y_{ijk}^2 - CT = \sum_{i=1}^{4}\sum_{j=1}^{2}\sum_{k=1}^{480} y_{ijk} - CT$$
$$= T - CT = 400 - 41.667 = 358.333$$

$$S_A = \sum_{i=1}^{4}\frac{T_{i..}^2}{mr} - CT = \frac{1}{960}[60^2 + \cdots + 168^2] - 41.667 = 10.566$$

$$S_B = \sum_{j=1}^{2}\frac{T_{.j.}^2}{lr} - CT = \frac{1}{1920}[160^2 + 240^2] - 41.667 = 1.6663$$

$$S_{AB} = \sum_{i=1}^{4}\sum_{j=1}^{2}\frac{T_{ij.}^2}{r} - CT = \frac{20^2 + \cdots + 88^2}{480} - 41.667 = 12.533$$

$$S_{E_1} = S_{A\times B} = S_{AB} - S_A - S_B = 0.3007$$

$$S_{E_2} = S_T - S_{AB} = 358.333 - 12.533 = 345.8$$

정리된 분산분석표는 다음 <표 16>과 같다.

<표 16> 모의자료 분산분석표

출처	SS	ν	MS	F_0	$p-value$
A	10.566	3	3.522	39.03^{**}	< 0.01
B	1.6663	1	1.6663	18.47^{**}	< 0.50
$E_{1(A\times B)}$	0.3007	3	0.1002	1.11	
E_2	345.8	3832	0.09024		
T	358.333	3839			

우선 1차단위의 오차 $E_1(A\times B)$의 유의성 검정에서 $F_0 = 1.11$이므로 유의하지 않은 것으로 판단하여 풀링하기로 결정하였다면 분산분석표는 <표 17>과 같이 다시 작성된다.

<표 17> 풀링된 분산분석표

출처	SS	ν	MS	F_0	$p-value$
A	10.566	3	3.522	39.03^{**}	< 0.01
B	1.6663	1	1.6663	18.47^{**}	< 0.50
E	346.1	3835	0.0902		
T	358.33	3839			

검정결과 A, B인자 모두 유의하다고 판단된다. 따라서 교호작용은 무시되며 모형은

$$y_{ijk} = \mu + a_i + b_j + \varepsilon_{ijk} \qquad y_{ijk} = \begin{cases} 0 : \text{적합} \\ 1 : \text{부적합} \end{cases} \qquad (30)$$
$$i = 1, 2, 3, 4 \quad j = 1, 2 \quad k = 1, \cdots, 480$$

와 같이 축소된다.

3.5.2 부적합률 추정

개선된 조업조건은 유일하게 A_3B_1이다. 교호작용이 무시되는 경우이므로 비율의 추정은

$$\tilde{p}(A_3B_1) = \tilde{p}(A_3) + \tilde{p}(B_1) - \tilde{p}$$
$$= 4.58 + 8.33 - \left(\frac{100}{960}\times 100\right) = 2.493(\%)$$

유효반복수는 다음과 같이 계산된다.

$$n_e = \frac{lmr}{l+m-1} = \frac{4\times 2\times 480}{4+2-1} = 768$$

90% 신뢰구간 폭은 다음과 같다.

$$t_{\alpha/2}(\nu_E)\sqrt{\frac{MS_E}{n_e}} = t_{0.05}(3835)\sqrt{\frac{0.0906}{768}}$$
$$= 1.645\times 0.01086 = 0.01786$$

분산분석에서는 각 수준조합에서의 분산이 동일하다는 가정 하에서 모형을 수립하고 분석하기 때문에 신뢰구간의 폭은 동일하게 적용되어야 한다. 따라서 $p(A_3B_1)$에 대한 90% 신뢰구간은

$$p(A_3B_1) : 2.50 \pm 1.786(\%) = (0.714\%, 4.286\%)$$

시행횟수를 4배로 증가시켰기 때문에 신뢰구간의 폭이 절반으로 줄어드는 것은 당연한 결과이다. 마찬가지 계산방식으로 현재의 조업조건 A_1B_1에서의 $p(A_1B_1)$에 대한 90% 신뢰구간은

$$p(A_1B_1) : 4.167 \pm 1.786(\%) = (2.381\%, 5.953\%)$$

신뢰구간이 상당부분 겹쳐서 나타나고 있어 변별력이 있다고 할 수 없는 것으로 보인다.

이를 좀 더 극명하게 나타내기 위해서는 차이에 대한 신뢰구간을 구해 볼 수 있다. 분산분석에서는 표본분포를 대칭적인 정규분포로 가정하고 있기 때문에 이 두개의 신뢰구간을 합성하여 쉽게 차이에 대한 신뢰구간을 구할 수 있다.

$$\begin{aligned} &\tilde{p}(A_1B_1) - \tilde{p}(A_3B_1) \\ &= \tilde{p}(A_1) + \tilde{p}(B_1) - \tilde{p} - \tilde{p}(A_3) - \tilde{p}(B_1) + \tilde{p} \\ &= \tilde{p}(A_1) - \tilde{p}(A_3) = \tilde{p_1} - \tilde{p_3} \end{aligned} \tag{31}$$

$$\tilde{p_1} - \tilde{p_3} = \frac{60}{960} - \frac{44}{960} = \frac{16}{960} = 0.01667 = 1.667(\%)$$

실제로 $\tilde{p}(A_1B_1) - \tilde{p}(A_3B_1)$에 대한 90% 신뢰구간은 다음과 같이 구해질 수 있다.

$$\begin{aligned} &(4.167 - 2.50) \pm \sqrt{2} \times 1.786 \\ &= 1.667 \pm 2.526 = (-0.859, 4.193) \end{aligned}$$

신뢰구간에 0이 포함되어 있어 두 모 비율이 서로 다르다는 근거로 채택하기 어렵다.

3.5.3 로지스틱 회귀의 적용

한편 로지스틱 회귀를 적용하여 볼 경우 적합된 로짓 연결함수는

$$\begin{aligned} g(\hat{p_{ij}}) &= \ln\left[\hat{p_{ij}}/(1-\hat{p_{ij}})\right] = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} \\ &= \alpha + \hat{\beta_2}x_2 + \hat{\beta_3}x_3 + \hat{\beta_4}x_4 + \hat{\gamma_2}x_6 + \hat{\delta_{42}}x_4x_6 \\ i &= 1, 2, 3, 4 \qquad j = 1, 2 \end{aligned} \tag{32}$$

와 같고 계수의 추정치는 다음과 같다.

$$\begin{aligned} &\hat{\alpha} = -3.1174,\ \hat{\beta_2} = 0.84473\ ,\hat{\beta_3} = -0.32975 \\ &\hat{\beta_4} = 1.5079,\ \hat{\gamma_2} = 0.70988,\ \hat{\delta_{42}} = -0.59437 \end{aligned}$$

<표 15>의 모의자료에 앞에서와 동일한 모형을 사용하여 동일한 방식으로 적합하여 본 결과 추가로 2개의 회귀계수 β_3와 δ_{42}가 신뢰수준 90%에서 거의 유의한 것으로 검출되었다. 유의한 것으로 판정된 계수가 더 많아졌을 뿐 만 아니라 계수의 추정치가 다소 변하였으며 신뢰구간이 더 좁아진다는 사실을 확인 할 수 있다.

<표 18> 모의자료 로지스틱 회귀 분석 표

예측						90% CI	
변수	계수	SE 계수	Z	P	승산비	하한	상한
상수	-3.11737	0.163701	-19.04	0.000			
x2							
1	0.844734	0.164477	5.14	0.000	2.33	1.78	3.05
x3							
1	-0.329747	0.204557	-1.61	0.107	0.72	0.51	1.01
x4							
1	1.50793	0.204446	7.38	0.000	4.52	3.23	6.32
x6							
1	0.709877	0.144857	4.90	0.000	2.03	1.60	2.58
x4x6							
1	-0.594365	0.223379	-2.66	0.008	0.55	0.38	0.80

이와 같이 이항 로지스틱 회귀모형에서는 시행횟수가 많으면 많을수록 신뢰구간이 점차 좁아져 유효성 있는 추정치(efficient estimate)를 구할 수 있으며 모형의 모수가 유의한 것으로 판명될 개연성이 높아진다는 사실을 확인 할 수 있다. 각 수준조합에서의 추정치와 신뢰구간을 모두 구한 결과는 <표 19>와 같다.

<표 19> 발생률 추정치와 신뢰구간

수준		발생률	추정치	90% 신뢰구간	
A	B	y_{ij}/n_{ij}	$\hat{p_{ij}}$	하한	상한
1	1	20/480	0.042397	0.032716	0.054780
1	2	40/480	0.082603	0.066438	0.102271
2	1	48/480	0.093415	0.076578	0.113499
2	2	80/480	0.173252	0.148904	0.200642
3	1	12/480	0.030855	0.023071	0.041155
3	2	32/480	0.060812	0.047219	0.077997
4	1	80/480	0.166667	0.140531	0.196552
4	2	88/480	0.183333	0.156045	0.214182

주) 현재 조업조건 A_1B_1, 개선된 조업조건 A_3B_1.

이와 같이 $p(A_3B_1)$에 대한 90% 신뢰구간은

$$p(A_3B_1) : (2.3071\%, 4.1155\%)$$

로서 분산분석에 의한 신뢰구간$(0.714\%, 4.286\%)$ 보다 더 좁게 나타나고 있다. 마찬가지 계산방식으로 현재의 조업조건 A_1B_1에서의 $p(A_1B_1)$에 대한 90% 신뢰구간은

$$p(A_1B_1) : (3.2716\%, 5.4780\%)$$

로서 분산분석에 의한 신뢰구간(2.381%, 5.953%)보다 더 개선되어 나타나고 있다. 따라서 분산분석법에 비하여 변별력이 더 우수하다. 하지만 이러한 해석은 분산분석의 관점에서의 해석이라고 할 수 있다.

로지스틱 회귀의 관점에서의 해석은 다소 다르게 진행되어야 한다. <표 18>에서 OR(odds ratio)과 90% 신뢰구간을 검토해 보면 OR추정치가 0.72이고 90% 신뢰구간이 (0.51, 1.01)인 것을 감안할 때 개선된 조업조건에서의 모비율 $p(A_3B_1) = p_{31}$이 더 작다는 사실을 거의 95% 신뢰할 수 있다. 그 이유는

$$OR = \frac{p_{31}/(1-p_{31})}{p_{11}/(1-p_{11})} \leq 1 \tag{33}$$

와 같이 OR(odds ratio)이 1보다 작거나 같다면

$$\begin{gathered} p_{31}/(1-p_{31}) \leq p_{11}/(1-p_{11}) \\ p_{31} - p_{31}p_{11} \leq p_{11} - p_{31}p_{11} \\ p_{31} \leq p_{11} \end{gathered} \tag{34}$$

와 같이 개선된 조업조건에서의 모비율(p_{31})이 현재의 경우의 모비율(p_{11})보다 작거나 같다는 의미로 해석이 가능하며 90% 양측 신뢰구간의 상한은 95% 좌측 신뢰구간의 상한에 해당한다.

이와 같이 로지스틱 회귀에 의한 해석은 분산분석방법론과 다소 다르기도 하거니와 변별력 면에서 더욱 우수하다고 할 수 있다.

4. 결론 및 제언

본 연구에서는 로지스틱 회귀 방식을 사용하여 백분비 또는 발생비율의 추정 시 발생되는 여러 가지 문제점들을 해결하려고 시도하였다.

정리된 계수치 자료만 있다면 최적 조업조건을 알아내는 것은 일견 쉬운 일로 보인다. 하지만 현재의 조업조건에 비해 통계적으로 유의한지(significant)를 알아야만 진정한 의미의 최적 조업조건을 찾았다고 말할 수 있다.

분산분석법은 모형의 기본 가정이 정규성, 독립성, 등분산성의 가정 하에서만 성립되는 모형이다. 더욱이 비율의 구간추정은 비대칭적으로 이루어져야 하는데[1] 정규 근사 한다는 것은 대칭적인 분포를 가정하고 구간추정을 하기 때문에 여러 가지 문제가 발생될 수 있다. 또한 분산분석이나 회귀분석의 선형모형에서는 가법성이라는 특성이 만족되어야 하는데 사건 발생비율이 0에 가깝거나 1에 매우 가까울 경우 성립되기가 어렵다.

이항 로지스틱 회귀 방식을 사용하는 경우 몇 가지 장점이 있는데 (1) 시행횟수가 반영된 추정결과가 구해지므로 시행횟수가 많을수록 정밀도가 높고 통계적으로 변별력 있는 바람직한 추정치를 얻을 수 있다. 또한 (2) 모수의 유의성을 별도의 검정과정 없이 OR의 구간추정만으로 검정할 수 있다. 단점이 있다면 (1) 로지스틱 회귀모형분석과정이 관례적인 방식과 다르기 때문에 분석 시 더 많은 주의력을 요한다. (2) 모수의 추정과정이 일반화 추정식(GEE)을 사용하여야 하므로 수작업계산은 불가능하다. 하지만 이러한 단점은 상업화된 통계패키지 예를 들어 SAS PROC LOGISTIC[4]이나 미니탭의 로지스틱 회귀 분석절차[7]를 이용한다면 쉽게 극복해 낼 수 있다.

앞으로 이러한 방법론을 확대 적용하여 계수치 자료 또는 여러 가지 범주형(categorical)자료 에도 실험계획법 또는 회귀분석에서와 같이 다양한 자료구조 모형에 적용하고 분석할 수 있을 것으로 기대된다.

참고문헌

[1] 류제복, 이승주; 낮은 이항 비율에 대한 신뢰구간, 응용통계연구, 19(2) : 217-230, 2006.

[2] 박성현; 현대실험계획법, 민영사, 2007.

[3] 박성현; 회귀분석, 제3판, 민영사, 1999.

[4] Allison, P. D.; *Logistic Regression Using the SAS System-Theory and App*, SAS, 1999.

[5] Dobson, A. J.; *An Introduction to eneralized Linear Models*, Chapman and Hall/CNC, 2001.

[6] Kleinbaum, D. G. and Klein, M.; *Logistic Regression : A Self Learming Text*, 3rd Edition Springer, 2010.

[7] Minitab; *Minitab Manual*, 2011.

[8] Montgomery, D. C., Peck, E. A., and Vining, G. G.; *Introduction to Linear Regression Analysis, 4th Edition*, 2006.

[9] Ross, P. J.; *Taguchi Techniques for Quality Engineering*, McGraw Hill, 1989.

[10] Sloan, D. and Morgan, S. P.; "*An Introduction to Categorical Data Analysis*," *Annual Review of Sociology*, 22 : 351-375, 1996.

[11] Strokes, M. E. Davis, C. S., and Koch, G. G., *Categorical Data analysis Using The SAS System*, 2nd Ed., 2000.