14th Week Assignment

AUTHOR
서강대학교 20121802 김재현

1. 로지스틱 회귀분석이 확률을 이용하지 않고, 로그오즈를 이용하여 확률을 표현하는 이유에 대하여 논의하시오.

회귀분석은 종속변수가 실수영역에 존재한다. 그러나 로지스틱 회귀분석의 종속변수인 확률은 0과 1 사이에만 존재한다. 오즈는 양수 영역에만 존재하며 그래프를 그려보면 대칭성도 없고 0.9 근처에서 급격히 숫자가 커지는 등 회귀분석에 적합하지 않다.

로그오즈는 양수와 음수 모두 존재하여 실수영역을 모두 커버한다. 또한 0을 기준으로 대칭을 이루는 그래프를 그린다. 그래프는 매우 부드러운 곡선을 갖고 있기 때문에 로그오즈가 회귀분석에 적합하여 로지스틱 회귀분석시 로그오즈를 이용하여 확률을 표현한다.

---

2. 어느 자동차회사의 마케팅 부서에서는 로지스틱 회귀분석을 이용하여 자동차에 대한 구매의도를 파악할 수 있는지를 알아보기 위하여 33명의 소비자를 대상으로 예비조사를 실시하였다. 랜덤으로 선택된 33명의 소비자를 대상으로 연간 소득과 조사 당시 보유하고 있는 차의 보유기간을 조사하였다. 이러한 조사가 끝난 후, 1년 뒤에 추적조사(follow up study)를 실시하여 신차의 구입여부를 조사하였더니, 아래와 같은 결과가 나타났다.

소비자 id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
구입여부 (1=구입)	0	0	1	0	0	1	1	1	0	1	1	0	0	1	0	0	1
연간소득 (백만원)	32	45	60	53	25	68	82	38	67	92	72	21	26	40	33	45	61
보유차량의 기간	3	2	2	1	4	1	2	5	2	2	3	5	3	4	3	1	2
소비자 id	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
구입여부 (1=구입)	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	
연간소득 (백만원)	16	18	22	27	35	40	10	24	15	23	19	22	61	21	32	17	
보유차량의 기간	3	4	6	3	3	3	4	3	4	3	5	2	2	3	5	1	

(1) 로지스틱 회귀분석을 실시하여 결과에 대하여 논의하시오.

```
# 데이터 불러오기
구입 = c(0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0
소득 = c(32, 45, 60, 53, 25, 68, 82, 38, 67, 92, 72, 21, 26, 40, 33, 45, 61, 16, 18, 22, 27, 35, 40, 10, 2
기간 = c(3, 2, 2, 1, 4, 1, 2, 5, 2, 2, 3, 5, 3, 4, 3, 1, 2, 3, 4, 6, 3, 3, 3, 4, 3, 4, 3, 5, 2, 2, 3, 5, 1
```

```
# 구입 변수는 factor로 변경
구입 = as.factor(구입)
```

```
# 최대 우도 추정을 통해 최적의 로지스틱 회귀 모형 추정
fullmodel = glm(구입 ~ 소득 + 기간 + 소득*기간, family=binomial())
nullmodel = glm(구입 ~ 1, family=binomial())
scope2 = formula(~ 소득 + 기간 + 소득*기간)
```

```
# Full model -> null model로 가면서 최적 로지스틱 회귀모형 추정
step(fullmodel, direction="backward")
```

```
Start:  AIC=43.4
구입 ~ 소득 + 기간 + 소득 * 기간
```

```
           Df Deviance    AIC
- 소득:기간  1   36.690 42.690
<none>           35.404 43.404

Step:  AIC=42.69
구입 ~ 소득 + 기간

       Df Deviance    AIC
<none>      36.690 42.690
- 기간  1   39.305 43.305
- 소득  1   44.987 48.987

Call:  glm(formula = 구입 ~ 소득 + 기간, family = binomial())

Coefficients:
(Intercept)         소득         기간
   -4.73931      0.06773      0.59863

Degrees of Freedom: 32 Total (i.e. Null);  30 Residual
Null Deviance:      44.99
Residual Deviance: 36.69    AIC: 42.69
```

```r
# null model에서 backward, forward 모두 사용하여 최적 모형 찾기
step(nullmodel, scope2, direction="both")
```

```
Start:  AIC=46.99
구입 ~ 1

       Df Deviance    AIC
+ 소득  1   39.305 43.305
<none>      44.987 46.987
+ 기간  1   44.987 48.987

Step:  AIC=43.3
구입 ~ 소득

       Df Deviance    AIC
+ 기간  1   36.690 42.690
<none>      39.305 43.305
- 소득  1   44.987 46.987

Step:  AIC=42.69
구입 ~ 소득 + 기간

           Df Deviance    AIC
<none>          36.690 42.690
- 기간       1   39.305 43.305
+ 소득:기간  1   35.404 43.404
- 소득       1   44.987 48.987


Call:  glm(formula = 구입 ~ 소득 + 기간, family = binomial())

Coefficients:
(Intercept)         소득         기간
   -4.73931      0.06773      0.59863

Degrees of Freedom: 32 Total (i.e. Null);  30 Residual
Null Deviance:      44.99
Residual Deviance: 36.69    AIC: 42.69
```

```r
# null model -> full model로 가면서 최적 모형 찾기
step(nullmodel, scope2, direction="forward")
```

```
Start:  AIC=46.99
구입 ~ 1

       Df Deviance    AIC
+ 소득  1   39.305 43.305
<none>      44.987 46.987
+ 기간  1   44.987 48.987

Step:  AIC=43.3
구입 ~ 소득

       Df Deviance    AIC
+ 기간  1   36.690 42.690
<none>      39.305 43.305

Step:  AIC=42.69
구입 ~ 소득 + 기간

            Df Deviance    AIC
<none>           36.690 42.690
+ 소득:기간  1   35.404 43.404


Call:  glm(formula = 구입 ~ 소득 + 기간, family = binomial())

Coefficients:
(Intercept)         소득         기간
   -4.73931      0.06773      0.59863

Degrees of Freedom: 32 Total (i.e. Null);  30 Residual
Null Deviance:      44.99
Residual Deviance: 36.69    AIC: 42.69
```

```r
# 최적 모형과 full model의 적합도 비교(카이제곱검정 분산분석)
M1 = glm(formula = 구입 ~ 소득 + 기간, family = binomial())
anova(M1, fullmodel, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: 구입 ~ 소득 + 기간
Model 2: 구입 ~ 소득 + 기간 + 소득 * 기간
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        30     36.690
2        29     35.404  1   1.2855   0.2569
```

> ⓘ Note
>
> p-value가 0.2569로 최적모형과 full model의 차이인 상호작용항(소득*기간)이 통계적으로 유의하지 않다(유의수준 5% 기준)는 것을 보여준다.

(2) 연간 5천만원의 소득이 있는 가정에서 현재 3년이 지난 차를 보유하고 있다면, 다음 해에 자동차를 구입할 확률이 어떻게 되는지를 계산하시오.

```r
# 소득 50(백만원), 기간 3(년)일 때 다음해 신차 구입 확률 - predict

predict(M1, list(소득=50, 기간 = 3), type="response")
```

```
        1
0.6090245
```

> ⓘ Note
>
> 약 60.9%의 확률로 다음해에 자동차를 구입할 것이다.

(3) 어느 가정이 다른 가정과 비교하여, 2천만원 소득이 높고, 2년 오래된 차를 보유하고 있다면, 신차를 구입할 오즈는 어떻게 변하며, 확률은 어떻게 변하는지에 대하여 신뢰구간을 이용하여 설명하시오.

```
# 최적 모형 확인

M1 = glm(formula = 구입 ~ 소득 + 기간, family = binomial())
summary(M1)
```

```
Call:
glm(formula = 구입 ~ 소득 + 기간, family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73931    2.10195  -2.255   0.0242 *
소득          0.06773    0.02806   2.414   0.0158 *
기간          0.59863    0.39007   1.535   0.1249
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 36.690  on 30  degrees of freedom
AIC: 42.69

Number of Fisher Scoring iterations: 4
```

```
# 최적모형의 coefficient를 이용하여 odds ratio 구하기

exp(coef(M1))
```

```
(Intercept)        소득        기간
0.008744682 1.070079093 1.819627221
```

> ⓘ Note
>
> 소득이 백만원 증가할때 신차 구입 오즈는 약 1.07배가 된다. 자동차 보유 기간이 1년 증가할 때 마다 신차 구입 오즈는 약 1.82배가 된다.

2천만원 소득이 높고, 2년 오래된 차를 보유하고 있다면

```
# 2천만원 소득이 높고, 2년 오래된 차를 보유하는 경우의 odds ratio 계산
(1.07)^20*(1.82)^2
```

```
[1] 12.81794
```

> ⓘ Note
>
> 오즈는 12.81배 커진다.

예를 들어, 연간 5천만원의 소득이 있는 가정에서 현재 3년이 지난 차를 보유하고 있는 가정이 신차를 구입할 확률은 0.6090245 이며 이에 대한 오즈는 아래와 같다.

```
# 연간 5천만원의 소득이 있는 가정에서 현재 3년이 지난 차를 보유하고 있는 가정의 오즈 계산
logit = predict(M1, list(소득=50, 기간 = 3), type="link")
exp(logit)
```

```
       1
1.557705
```

그러므로, 비교 가정보다 2천만원 소득이 높고, 2년 오래된 차를 보유하고 있는(연 7천만원 소득, 5년이 지난 차를 보유하고 있는) 가정이 신차를 구매할 오즈는 아래와 같다.

```
# 연 7천만원 소득, 5년이 지난 차를 보유하고 있는 가정의 오즈 계산

1.557705*12.81794
```

```
[1] 19.96657
```

이를 확률로 변환하면,

```
# 오즈를 확률로 변환

19.96657/(1+19.96657)
```

```
[1] 0.952305
```

> ⓘ Note
> 신차를 구입할 확률은 0.6090245 에서 0.952305으로 증가했다.

```
# 소득과 기간의 오즈 신뢰구간 확인
exp(confint(M1, c("소득","기간"))) # Approximation of 95% CI for Odds Ratio
```

```
Waiting for profiling to be done...

        2.5 %   97.5 %
소득 1.019694 1.140854
기간 0.884999 4.250058
```

> ⓘ Note
> 신뢰구간을 확인해보면 소득의 오즈의 신뢰구간은 1을 포함하고 있지 않기 때문에, 소득수준에 따라 신차를 구매할 확률에 확실히 차이가 있다고 볼 수 있다.
>
> 그러나, 기간의 오즈의 신뢰구간은 1을 포함하고 있으므로, 기간이 신차 구매 확률에 끼치는 영향은 있을 수도 있고, 없을 수도 있다

(4) 보유차량 기간을 독립변수로 포함시킨 로지스틱 회귀모형과 포함시키지 않은 회귀모형을 비교하여 분석하고, 결과에 대하여 논의하시오.

```
# 기간을 독립변수로 포함시킨 로지스틱 회귀모형과 제외한 모형의 적합도 비교(카이제곱검정 분산분석)


M1 = glm(formula = 구입 ~ 소득 + 기간, family = binomial())
M1_updated = update(M1, formula = . ~ . - 기간) # 기존 모델 M1에서 '기간' 변수를 제거

anova(M1_updated, M1, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: 구입 ~ 소득
Model 2: 구입 ~ 소득 + 기간
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        31     39.305
2        30     36.690  1   2.6149   0.1059
```

> ⓘ Note
> p-value가 0.1059로 기간항이 통계적으로 유의하지 않다(유의수준 5% 기준)는 것을 보여준다.

(5) 연간소득과 보유차량의 기간 사이에 존재할 수 있는 상호작용효과는 어떠한 의미를 갖는지에 대하여 설명하시오.

연간소득이 많지만 보유 차량의 기간이 긴 경우, 해당 가구는 절약하는 경향이 있다고 볼 수 있다. 이러한 가구는 신차가 아닌 중고차를 구매할 가능성이 더 커보인다. 때문에 연간소득이 많으면서 보유 차량의 기간이 길면, 차를 바꿀 수 있을 만큼 소득이 충분하고 차를 바꿀 동기가 충분하기 때문에 신차를 구매할 확률이 매우 높아야 하지만, 절약하는 경향이 있는 가구들이 데이터에 포함되어 연간소득이 많고 보유 차량의 기간이 길지

만 신차를 구매할 확률을 낮춘다. 이러한 상호작용으로 인해 기간 독립변수가 신차를 구매할 확률에 미치는 영향이 통계적으로 유의하지 않게 나타나는 것으로 추측한다.

(6) 신차의 구입여부를 예측하기 위하여, 어떠한 다른 독립변수가 사용될 수 있는지에 대하여 논의하시오.

소득, 차량 보유기간과 더불어, 이전까지 신차를 구매했던 이력(수)를 독립변수로 사용하면 신차를 구매할 확률을 예측하는데에 도움을 줄 것이라 생각한다. 앞서 (5)에서 밝힌바와 같이 절약하는 가구는 신차를 구매하지 않고 중고차를 구매하는 경향이 있을 것이라 예상한다. 때문에, 신차를 구매했던 이력(수) 독립변수가 있다면, 신차 대신 중고차를 구매하는 가구를 분리할 수 있을 것으로 생각한다.