통계자료분석 12주차 과제

20191245 노유정

<정답>

1번

실험자가 분류에 필요한 다른 변수들을 고려하여 판단하게 하기 위하여 확률을 제시한다.

2번

일반적인 회귀분석은 실수 전체 범위의 값을 가지는데 확률은 0~1 사이의 값을 가지므로 바로 적용할 수 없어 이를 로그오즈로 변환하여 적용한다.

3번

J형태라면 확률 0.5를 기준으로 대칭되지 않기 때문에 로지스틱 회귀분석에 적용할 수 없다.

4번-1

신차 구입 여부는 연간소득과 보유차량의 기간의 영향을 받음을 알 수 있다. 소득에 비해 기간이 신뢰도가 떨어진다.

4번-2

약 60%

4번-3

오즈 12.7 증가하여 신차구입확률도 증가한다.

4번-4

차량 보유기간을 변수로 포함시켰을 때의 AIC값이 더 작으므로 기간을 변수로 포함시키는 것이 더 적합한 분석모형이 된다.

4번-5

자동차는 사치품의 특성을 강하게 띄기 때문에 경제적 여유가 있는 고소득자의 경우, 그렇지 않은 사람들보다 자동차를 더 자주 바꿀 것이라고 예측할 수 있다. 분석 결과 연봉이 3천만원 늘어날 때마다 차량 보유기간이 1년 줄어드는 것으로 관측되었다.

4번-6

집에서 회사까지의 거리, 연간 장거리 출장 횟수 등이 사용될 수 있다.

<해설>

1. 두 모집단을 분류하는 로지스틱 회귀분석에서 결과가 0이나 1로 나타나지 않고, 0에서 1사이의 확률로 표현하는 이유에 대하여 논의하시오.

회귀분석으로 바로 결과를 내는 것이 아닌, 실험자가 분류에 필요한 다른 중요 변수들을 같이 고려하여 판단하기 위하여 확률을 제시한다.

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

확률은 0에서 1 사이의 값을 지녀 실수 전체 범위의 값을 갖는 일반적인 회귀분석에 바로 적용할 수 없으므로, 마찬가지로 실수 전체 값을 갖는 로그오즈로 변환하여 사용한다.

3. 확률과 독립변수의 관계가 실제로 S 모양의 형태가 아니라 J 모양의 형태를 갖는 경우에도 로지스틱 회귀분석을 적용할 수 있는지에 대하여 논의하시오.

확률 0.5를 기준으로 대칭되지 않기 때문에 로그오즈로 표현하였을 때 0을 기준으로 대칭되지 않아 로지스틱 회귀분석에 적용할 수 없다.

4. 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

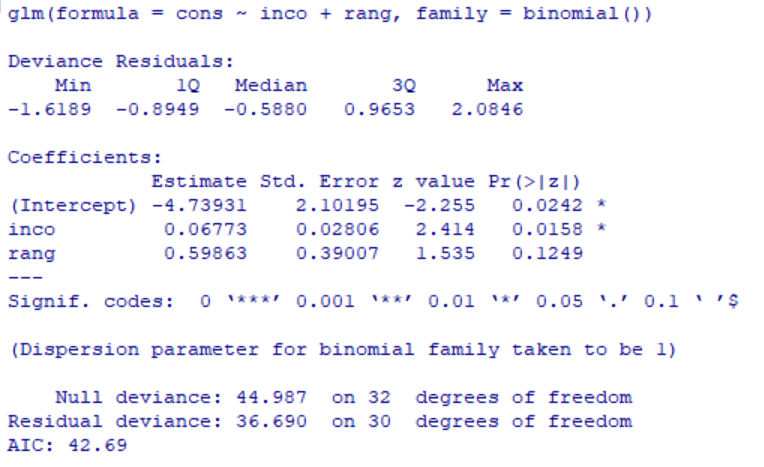
소비자 id             18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33

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.

신차 구입 여부는 연간소득과 보유차량의 기간의 영향을 받음을 알 수 있다. 소득은 p값이 0.0158, 기간은 0.1249로 기간의 신뢰도가 떨어지는데, 이는 자료에서 소득은 범위가 넓으며 단위도 자세한데 기간은 연단위로 끊어 1~6의 정수값만 가지며 1,2년 차이가 그리 크지 않기 때문에 그런 것으로 예측할 수 있다. 기간을 개월수로 하면 신뢰도를 조금 더 높일 수 있을 것이다.

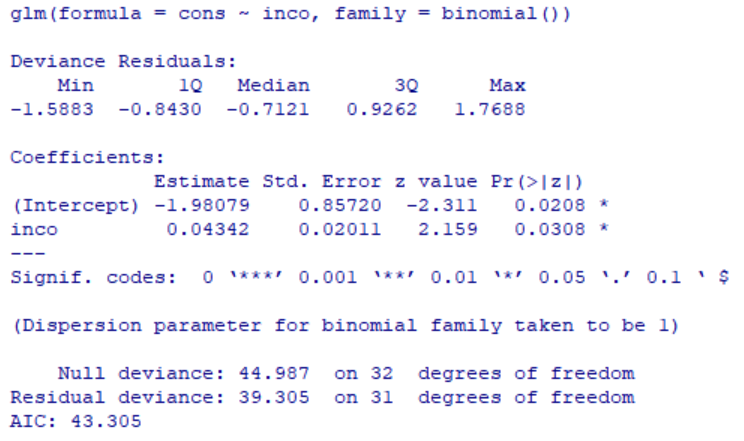
(2) 연간 5천만원의 소득이 있는 가정에서 현재 3년이 지난 차를 보유하고 있다면, 다음 해에 자동차를 구입할 확률이 어떻게 되는지를 계산하시오.

-4.74 + 0.067\*50 + 0.60\*3=0.41 , exp(0.38)=1.507 ,p= 0.601, 약 60%

(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.

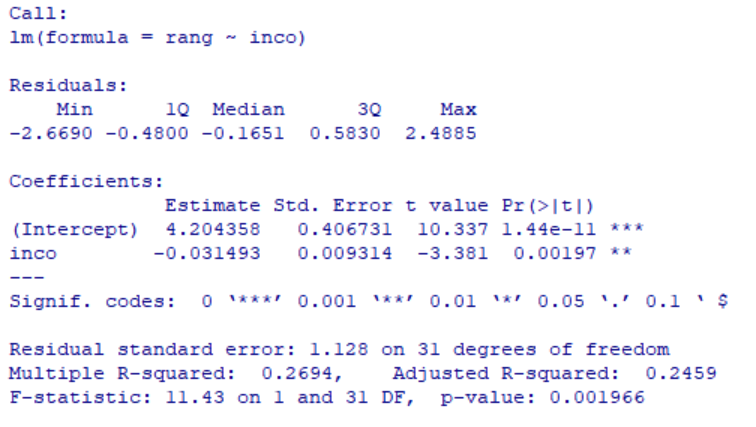
0.067\*20+0.60\*2=2.54, exp(2.54)=12.68로 오즈는 12.7 증가한다. 오즈가 증가하므로 확률도 증가한다.

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

전자의 AIC값은 42.69, 후자는 43.305로 차량 보유기간을 변수로 포함시켰을 때의 AIC값이 더 작으므로 기간을 변수로 포함시키는 것이 더 적합한 분석모형이 된다.

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

자동차는 소모품이면서 사치품의 특성을 강하게 띄기 때문에 경제적 여유가 있는 고소득자의 경우, 그렇지 않은 사람들보다 자동차를 더 자주 바꿀 것이라고 예측할 수 있다.

실제로 기간을 종속변수, 소득을 독립변수로 두어 선형회귀분석을 진행한 결과 약 99.9의 신뢰도로 미미하지만 음의상관관계가 있다고 판명되었다. 회귀계수는 -0.03으로 4.2년을 기준으로 소득이 약 3천만원 늘어날 때마다 보유기간이 1년 줄어든다고 분석된다.

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

집에서 회사까지의 거리, 연간 장거리 출장 횟수 등이 사용될 수 있다.