Untitled

SeungGyu Bak

2022-10-12

## 5.1 제5강 과제에서 분석한 참게자료(crabs)에 대해 부수체의 확률에 대한 무게와 너비를 설명변수로 하는 모형을 적합하여라.

dt1 <- read.table('Crabs.txt', header=TRUE)  
fit <- glm(y~weight+width, family = binomial, data = dt1)  
summary(fit)

##   
## Call:  
## glm(formula = y ~ weight + width, family = binomial, data = dt1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1127 -1.0344 0.5304 0.9006 1.7207   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.3547 3.5280 -2.652 0.00801 \*\*  
## weight 0.8338 0.6716 1.241 0.21445   
## width 0.3068 0.1819 1.686 0.09177 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 192.89 on 170 degrees of freedom  
## AIC: 198.89  
##   
## Number of Fisher Scoring iterations: 4

1. 가설 에 대한 가능도비 검정을 실시하고 그 결과를 해석하여라.

Y ~ weight + width 로지스틱 회귀모형을 적합한 결과, null 모형의 이탈도는 225.76(df=172)이고 적합한 모형의 이탈도는 192.89(df=170)으로 나타났다. 가능도비 검정을 위한 통계량은 로 인 카이제곱 분포를 따른다. 해당 통계량의 p값은 > 0.001로 계산된다. 따라서 모든 변수에 대한 회귀계수가 0이라는 귀무가설은 유의수준 0.05에서 기각된다.

1. 각 변수의 부분효과에 대하여 별도의 가능도비 검정을 실시하여라. 왜 (a)의 검정에서는 매우 유의하게 나오지만 각각의 검정에서는 그렇지 않은가?

로지스틱 회귀분석 결과, 중량과 너비 각각의 왈드 검정 p값은 모두 유의수준 0.05에서 귀무가설을 기각할 수 없는 것으로 나타난다. 각 변수에 대한 가능도비 검정 결과는 다음과 같다.

Anova(fit)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y  
## LR Chisq Df Pr(>Chisq)   
## weight 1.5608 1 0.21155   
## width 2.8453 1 0.09164 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

전체 모형의 p값은 낮게 나오는 반면 개별 변수값의 검정결과는 유의수준에 미치지 못하는 사례는 다중공선성(multicollinearity) 때문에 나타난다. 투구게 사례의 경우 중량과 너비 각각은 모두 부수체의 수 또는 부수체 보유 확률을 잘 설명해주는 변수이다. 하지만 투구게 암컷의 개체연령이 늘어나는 경우 중량이 늘어나고 너비도 커진다. 즉, 개체연령이라는 제3의 변수가 중량과 너비에 직접 영향을 미치기 때문에, 중량과 너비는 독립성이 보장되지 않는다. 이러한 점은 중량과 너비의 상관계수가 1에 가까운 값을 가지는 점을 통해서도 확인할 수 있다.

cor(dt1[,4:5])

## weight width  
## weight 1.0000000 0.8868715  
## width 0.8868715 1.0000000

1. 무게와 등뼈 상태 및 색깔이 잠재적 설명변수(potential explanatory variables)일 때 목적적 선택(purposeful selection) 또는 AIC를 사용하여 모델을 구축하라.

<1단계 : 잠재변수 중 변수 모형에 포함될 변수 선택>

|  | Deviance | df | AIC | LR Stat. | LR df |
| --- | --- | --- | --- | --- | --- |
| None | 239.8289 | 173 | 239.8289 | NA | NA |
| Weight | 195.7371 | 171 | 199.7371 | 30.021373 | 1 |
| Spine | 225.6243 | 171 | 229.6243 | 0.134229 | 1 |
| Color | 213.2979 | 171 | 217.2979 | 12.460655 | 1 |

분석의 대상이 된 중량, 등뼈, 색깔을 대상으로 각 변수를 융리변수로 하는 모형을 null모형과 비교한 결과, 중량과 색깔은 유의수준 0.2(임의로 설정한 값) 수준에서 LR 검정을 통과하지만, 등뼈는 유의수준을 충족하지 못한다. 따라서 등뼈는 잠재변수에서 제외하고 중량과 색깔만을 대상으로 다음 단계를 진행한다.

<2단계 : 후진소거>

|  | Deviance | df | AIC | LR Stat. | LR df |
| --- | --- | --- | --- | --- | --- |
| C + W | 190.2688 | 170 | 196.2688 | 35.489742 | 2 |
| Weight | 195.7371 | 171 | 199.7371 | 5.468368 | 1 |
| Color | 213.2979 | 171 | 217.2979 | 23.029087 | 1 |

C+W 모형을 기준으로 각 변수에 대해 후진소거한 결과는 위와 같다. 중량과 색깔을 각각 제거한 모형과 비교한 결과 C+W 모형은 유의성을 가지는 것으로 나타났다. 따라서 후진소거없이 C+W 모형을 선택한다.

<3단계 : 잠재변수 추가>

|  | Deviance | df | AIC | LR Stat. | LR df |
| --- | --- | --- | --- | --- | --- |
| C + W | 190.2688 | 170 | 196.2688 | 35.489742 | 2 |
| C + W + S | 188.7169 | 169 | 196.7169 | 1.551926 | 1 |

1단계에서 제거한 등뼈 변수를 추가한 모형과 비교한 결과, 통계적 유의성이 없는 것으로 확인되었다. 따라서 변수 추가 없이 C + W 모형을 유지한다.

<4단계 : 교호작용 고려>

|  | Deviance | df | AIC | LR Stat. | LR df |
| --- | --- | --- | --- | --- | --- |
| C + W | 190.2688 | 170 | 196.2688 | 35.4897416 | 2 |
| C + W + CW | 190.1897 | 169 | 198.1897 | 0.0791072 | 1 |

각 변수 간의 교호작용을 고려한 결과, 교호작용을 고려하지 않은 모형에 비해 통계적인 유의성이 없는 것으로 확인되었다.

따라서 교호작용에 대한 고려없이 중량과 색깔만을 포함한 모형을 최종선택할 수 있다.

## 5.5 다음은 Myers-Briggs 적성검사의 네 가지 척도 자료이다.

dt5 <- read.table('MBTI.txt', sep = ',', header = TRUE)  
dt5

## EI SN TF JP smoke drink n  
## 1 e s t j 13 10 77  
## 2 e s t p 11 8 42  
## 3 e s f j 16 5 106  
## 4 e s f p 19 7 79  
## 5 e n t j 6 3 23  
## 6 e n t p 4 2 18  
## 7 e n f j 6 4 31  
## 8 e n f p 23 15 80  
## 9 i s t j 32 17 140  
## 10 i s t p 9 3 52  
## 11 i s f j 34 6 138  
## 12 i s f p 29 4 106  
## 13 i n t j 4 1 13  
## 14 i n t p 9 5 35  
## 15 i n f j 4 1 31  
## 16 i n f p 22 6 79

데이터 링크에 들어가면 16개의 예측변수의 조합이 있고, ESTJ형 예를 들어 설명하면, 77명 중에서 10명이 음주 빈도가 높은 것으로 보고했다.

<표 5.6>는 이 네 가지 척도를 예측변수로 사용하여 연구 참여자가 술을 자주 마시는지 여부를 예측하기 위해 모형을 적합한 SAS 출력 결과이다.

1. 적합도 검정을 수행하고 해석하라. 만약 예측변수를 제거하여 모형을 단순화하고자 한다면 어떤 변수를 제거할 것인가? 그 이유는 무엇인가?

모형 평가를 위한 적합도 검정은 주어진 데이터를 포화모형에 근거한다고 가정하고, 유추한 모형에서 누락된 변수의 회귀계수가 0인지를 측정하는 검정 방법이다. 귀무가설은 이 되며, 적합도 검정의 p값이 낮게 나오는 경우 모형에서 누락된 변수의 설명력이 높다는 결론을 내릴 수 있다.

주어진 분석 결과에서 잔차 이탈도(deviance)는 11.1591, 자유도는 11로 도출되었다. 이때 자유도는 이 된다. 해당 값을 바탕으로 한 카이제곱 검정 결과의 p값은 0.43으로, 귀무가설을 기각할 수 없는 것으로 나타난다. 따라서, 모형은 주어진 데이터를 잘 설명하는 것으로 결론내릴 수 있다.

각 변수를 하나씩 살펴보면, 95% 신뢰구간을 기준으로 EI, SN, TF 변수는 회귀계수가 0이 아니라고 결론내릴 수 있지만, JP 변수는 95% 신뢰구간이 (-0.6477, 0.2426)으로 회귀계수가 0이 아니라는 결론을 내릴 수 없다. 따라서, 모형을 단순화하기 위해 변수를 제거한다면 JP 변수를 제거하는 것이 타당하다.

1. 소프트웨어로부터 4개의 주효과와 6개의 교호작용항을 포함하는 모형은 AIC 값이 642.1이고, 4개의 이진변수의 주효과항만을 포함하는 모형은 637.5이고 예측변수가 없는 영모형의 경우는 648.8임을 알았다. 이 AIC 기준에 따르면 어떤 모형이 가장 선호되는가? AIC를 사용하는 타당성에 대해 설명하라.

아카이케 정보량 기준(AIC)은 모형을 적합하여 구한 값과 예상되는 값 간의 차이와 모형의 간결성을 기준으로 모형의 적합도를 측정하는 값이다. AIC가 낮을수록 좋은 모형으로 평가하며, 구하는 식은 다음과 같다.

모형이 주어진 데이터의 최대우도값을 정확하게 추정할수록(모형의 예측값과 관측값 간의 차이가 적을수록), 모형이 간결할수록(모수의 수가 적을수록) AIC 값은 작아진다. 일반적으로 기준치가 되는 AIC값은 없기 때문에 단일 모형의 AIC값이 적절한지를 평가하는 기준은 없으나 두 개 이상의 모형을 비교하는 경우 모형의 간결성과 정확성 간의 최적 균형점에 얼마나 근접해있느냐를 판단할 수 있는 통계량으로 활용된다.

주어진 사례에서 AIC를 기준으로 주효과항만을 포함하는 모형의 타당성이 가장 높은 것으로 나타났다. 즉, 변수가 많아지면 많아질수록 설명력이 높아지기 때문에. 와 같이 모형의 설명력만을 보여주는 통계량은 높아지지만 복잡한 모형의 경우 편차가 커지기 때문에 관심의 대상이 되는 변수가 가지는 효과를 명확하게 보여줄 수 없다는 한계를 가지고 있다. 반면, 변수가 적어지면 편향이 생겨서 모형의 정확도가 떨어질 수 있다. 이 두 가지 요인 간의 균형을 추구하는 통계량이 AIC이다. 주어진 사례에서 변수가 많은 모형의 경우 AIC가 높아지는데, 이는 변수를 추가해서 얻을 수 있는 정확도의 이득이 변수 추가로 인한 간결성의 손해보다 크지 않음을 뜻한다.

1. MBTI.dat에 대해 모형 구축 방법을 적용하여 음주 반응변수에 대한 모형을 선택하라.

< 후진소거법 >

MBTI 데이터에 대해 AIC값을 기준으로 하는 후진소거법을 이용하여 모형을 구축한 결과는 아래와 같다.

dt5\_1 <- dt5[,c(1, 2, 3, 4, 6)]  
fit5 <- glm(drink ~ ., family = poisson, data = dt5\_1)  
step5 <- step(fit5)

## Start: AIC=105.54  
## drink ~ EI + SN + TF + JP  
##   
## Df Deviance AIC  
## - TF 1 41.003 103.55  
## - JP 1 41.086 103.63  
## - EI 1 42.243 104.79  
## <none> 40.993 105.54  
## - SN 1 46.499 109.05  
##   
## Step: AIC=103.55  
## drink ~ EI + SN + JP  
##   
## Df Deviance AIC  
## - JP 1 41.096 101.64  
## - EI 1 42.253 102.80  
## <none> 41.003 103.55  
## - SN 1 46.509 107.06  
##   
## Step: AIC=101.64  
## drink ~ EI + SN  
##   
## Df Deviance AIC  
## - EI 1 42.346 100.89  
## <none> 41.096 101.64  
## - SN 1 46.602 105.15  
##   
## Step: AIC=100.89  
## drink ~ SN  
##   
## Df Deviance AIC  
## <none> 42.346 100.89  
## - SN 1 47.852 104.40

summary(step5)

##   
## Call:  
## glm(formula = drink ~ SN, family = poisson, data = dt5\_1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0462 -1.3839 -0.4326 0.2882 3.8141   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.5315 0.1644 9.316 <2e-16 \*\*\*  
## SNs 0.4834 0.2090 2.313 0.0207 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 47.852 on 15 degrees of freedom  
## Residual deviance: 42.346 on 14 degrees of freedom  
## AIC: 100.89  
##   
## Number of Fisher Scoring iterations: 5

< 목적적 선택 >

|  | Deviance | df | AIC | LR Stat. | LR df |
| --- | --- | --- | --- | --- | --- |
| None | 235.46390 | 16 | 290.0096 | NA | NA |
| EI | 46.60207 | 14 | 105.1477 | 1.2501102 | 1 |
| SN | 42.34629 | 14 | 100.8919 | 5.5058962 | 1 |
| TF | 47.84187 | 14 | 106.3875 | 0.0103095 | 1 |
| JP | 47.75938 | 14 | 106.3050 | 0.0927983 | 1 |

목적적 선택의 1단계인 잠재변수 선택을 위해 각 변수 1개씩만을 포함하는 모형을 적합한 결과값은 위 표와 같다. SN 이외 나머지 변수들은 0.02 수준에서 우도비 검정을 통과하지 못해서 SN만을 포함하는 모형을 선정한다. 2단계 수행을 위해 SN만을 포함하는 모형과 귀무모형을 비교한 결과 AIC값은 SN모형이 더 낮으므로 SN모형을 유지한다.

|  | Deviance | df | AIC | LR Stat. | LR df |
| --- | --- | --- | --- | --- | --- |
| SN | 42.34629 | 14 | 100.8919 | 5.5058962 | 1 |
| SN + EI | 41.09618 | 13 | 101.6418 | 1.2501102 | 1 |
| SN + TF | 42.33598 | 13 | 102.8816 | 0.0103095 | 1 |
| SN+ JP | 42.25349 | 13 | 102.7992 | 0.0927983 | 1 |

3단계로 SN모형에 변수를 추가한 분석 결과는 위 표와 같다. SN외에 어느 변수를 추가해도 유의미한 검정 결과가 도출되지 않는다. 따라서 SN모형을 유지한다. 단일 변수 모형이기 때문에 4단계인 교호작용은 고려할 여지가 없다.

따라서 최종 모형은 SN만을 변수로 포함하는 모형을 선택한다.