데이터 분석 방법론

이용희

2023-10-04, 오후 11 시

목차

그림 목록

표 목록

Preface

이 사이트는 데이터 분석 방법론 강의 온라인 강의 노트입니다.

1. 서론

1.1. 학습 내용

이 교과서는 다양한 형태를 가진 자료를 분석하는 통계적 방법들의 이론과 응용을 살펴보기 위한 것입니다.

이 교과서에서는 다음과 같은 주제를 다룰 것입니다.

- 교차표에서의 통계적 분석방법
- 범주형 자료와 발생횟수를 따르는 자료에 대한 모형 구축과 추론
- 일반화 선형모형에서의 추론
- 반복측정자료와 군집자료에 대한 분석 방법

1.2. R 언어

이 교과서에서는 통계 방법들의 실습을 위하여 R 프로그램을 사용합니다. R 프로그램이 익숙하지 않는 학생들은 R 프로그램에 대한 기초적인 내용을 먼저 숙지하느 것을 추천합니다. 참고로 저자의 R 기초 강의 사이트에서 R 프로그램에 대한 기초적인 내용을 배울 수 있습니다.

이 강의에서 사용하는 R 패키지는 다음과 같다.

1.3. 참고도서

- Faraway (2016)
- Agresti (2007)
- Agresti (2012)

1.4. 유용한 사이트

- Data sets for “An Introduction to Categorical Data Analysis”
- R codes for “An Introduction to Categorical Data Analysis”

Part I.

분할표의 분석

2. 연관성의 측도

2.1. 필요한 패키지

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
```

2.2. 이항변수

통계학에서 관측값은 값이 가지는 특성에 따라서 연속형 변수(continuous variable)과 범주형 변수(categorical variable)로 나눈다.

결과가 2개인 범주형 변수인 이항변수(binary variable)는 매우 중요한 역할을 한다. 그 이유는 두 개의 선택 중에서 하나를 선택해 야할 의사결정이 실제로 대부분을 차지하고 있기 때문이다.

예를 들어서 코로나 19에 감염된 환자가 병원에서 치료를 받고 있다고 가정해보자. 환자는 병원에서 여러 가지 검사를 수행하면서 다양한 자료를 수집한다. 예를 들어 환자는 수시로 체온을 재고 항체검사, 혈액검사 등을 받을 것이다. 다양한 검사 등에서 나온 자료는 연속형 또는 범주형 자료로 구성될 것이다.

하지만 의사가 가장 중요하게 결정할 사항은 환자가 계속 치료를 필요로 하는지 아닌지 결정해야 한다. 즉, 여러 가지 검사를 고려하여 최종적으로 의사는 환자가 더 치료가 필요한지 아닌 지를 결정해야 한다. 의사의 결정을 이항변수 Y로 다음과 같이 표현할 수 있다..

$$Y = \begin{cases} 1 & \text{if patient still needs treatment} \\ 0 & \text{if patient dose not need treatment any more (GO HOME!)} \end{cases}$$

실제 임상에서는 이러한 두 개의 가능한 선택 중에 하나를 선택하는 결정이 빈번하게 일어나며 이러한 결정은 대부분 중요한 임상적 결정이다. 예를 들어 다음과 같은 의사결정들은 이항변수로 표현할 수 있다.

- 환자는 약을 복용해야 하는가?
- 환자는 입원을 해야 하는가?
- 환자는 중환자실로 가야 하는가?
- 환자는 퇴원해도 되는가?

또는 환자의 상태(outcome)가 이항변수로 표현될 수 있다.

- 환자는 치료가 되었는가?
- 환자가 사망하였는가?

이제 코로나 19 치료제의 효과를 알아보기 위한 임상실험을 수행하는 경우를 생각해보자. 통상적으로 임상실험에서는 두 개의 집단을 비교하며 가장 많이 사용하는 두 개의 집단은 실제 치료(drug)를 받은 사람들과 위약(placebo)을 받은 사람들이다. 즉 치료를 받은 사람과 받지 않는 사람들의 효과를 비교하는 것이 임상실험의 목적이다. 이러한 경우 앞에서 논의한 의사 결정과 마찬가지로 한 환자가 받은 치료의 종류를 이항변수 X로 나타낼 수 있다.

$$X = \begin{cases} 1 & \text{if patient receives drug} \\ 0 & \text{if treatment receives placebo} \end{cases}$$

2.3. 분할표와 연관성의 측도

2.3.1. 분할표

이제 앞에서 말한 두 개의 변수 X 와 Y 의 관계에 대해서 생각해 보자. 실험에서 사람들은 코로나 19에 대한 치료약의 효과에 관심이 있다. 코로나 19 환자가 치료약을 처치 받으면 치료약을 이용하지 않는 환자보다 빨리 치료되거나 사망할 가능성이 낮은 지가 주요 관심사이다. 즉, 치료약이 환자의 회복 속도나 사망과 연관(association)이 있는지 알고 싶은 것이며, 특히 실험이 매우 정교하게 설계된 경우는 치료약이 환자의 회복이나 사망에 영향을 미치는 원인이 되는지(cause-effect relation) 파악하고 싶은 것이다.

- 먼저 코로나 19에 대한 치료약의 효과에 대한 임상실험에 n 명의 환자들이 실험에 참가 했다고 가정하자.
- 치료약이 효과가 있는지에 대한 결과(Y)는 치료를 시작하여 정해진 기간 내에 사망하였는지에 대한 사건으로 결정하였다.

$$Y = \begin{cases} 1 & \text{if patient is dead within D days} \\ 0 & \text{otherwise} \end{cases}$$

코로나 19에 대한 치료약의 효과에 대한 임상실험의 결과를 다음과 같은 **분할표(contingency table)**로 요약할 수 있다.

표 2.1.: 2×2 분할표

치료/결과	사망 ($Y=1$)	생존 ($Y=0$)	합계
위약 ($X=0$)	n_{11}	n_{12}	n_{1+}
치료약 ($X=1$)	n_{21}	n_{22}	n_{2+}
합계	n_{+1}	n_{+2}	n

많은 임상실험이나 의학연구의 결과들을 위와 같은 2×2 분할표로 요약할 수 있다. 이제 우리의 관심은 분할표를 통해서 임상실험의 결과를 어떻게 통계적으로 추론할 수 있는지이다.

> **노트**
>
> 분할표에서 연관성의 측도를 계산하는 경우 성공의 기준(이항변수로 표현하면 $Y=1$)에 따라서 계산을 수행해야 한다. 어떤 경우는 사망이나 악화와 같은 위험한 사건이 성공 사건이 될 수 있으며 어떤 경우는 생존이나 회복과 같은 좋은 사건이 성공이 될 수 있다.
> 또한 기준이 되는 그룹(이항변수 X)에 따라서 연관성의 측도 계산할 때 분자와 분모에 해당하는 그룹을 적절하게 선택해야 한다.
> 분할표에서 연관성의 측도를 계산하는 경우 분석의 의도와 목적에 맞게 성공 사건과 기준그룹을 정의하고 그에 따라서

연관성의 측도를 계산해야 한다.

2.3.2. 상대위험

2×2 분할표 표 2.1 에서 두 개의 처리군, 즉 치료약을 받은 집단과 위약을 받은 집단의 효과를 비교할 때 가장 많이 사용되는 측도(measure)는 **상대위험(relative risk, risk ratio, prevalnce ratio;RR)**이다.

주어진 집단의 위험율을 그 집단에 속한 환자의 수에서 사망한 사람의 비율이다. 분할표 표 2.1 에서 위약 집단의 위험율은 n_{11}/n_{1+} 이며 이는 치료를 받지 않는 경우에 나타나는 기준점인 위험율(baseline risk)을 의미한다. 치료약 집단의 위험율은 n_{21}/n_{2+} 이다. 통상적으로 위험율은 비율(proportion, percent)로 나타내며 발생률(rate,예를 들어 인구 1000명당 X명)로 나타내기도 한다.

상대위험은 두 위험율의 비율로서 다음과 같이 정의한다.

$$RR = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}} = \frac{\frac{n_{11}}{n_{11}+n_{12}}}{\frac{n_{21}}{n_{21}+n_{22}}} \tag{2.1}$$

상대위험이 1보다 크면 분자에 위치한 집단이 위험(위의 예제에서는 위험이 사망을 의미한다)에 처할 가능성이 분모에 위치한 집단보다 RR 배 높다는 것을 의미한다. 상대위험이 1이면 두 집단에 대한 위험이 동일하다는 것을 의미한다.

예를 들어 특정한 코로나 치료제의 효과를 실험하는 임상실험에서 다음과 결과를 얻었다.

표 2.2.: 코로나 치료제 실험 결과

치료/결과	사망 ($Y=1$)	생존 ($Y=0$)	합계
위약 ($X=0$)	10	1212	1222
치료약 ($X=1$)	5	2355	2360
합계	15	3567	3582

상대위험은 다음과 같이 계산된다.

$$RR = \frac{10/1222}{5/2360} = 3.8625 \approx 4$$

상대위험이 약 4 배란 의미는 치료약을 받은 집단보다 위약집단이 사망할 가능성이 약 4배 높다는 것이다.

i 노트

우리는 두 집단의 비율을 비교할 때 두 비율의 차이를 이용하는 방법을 자주 사용한다. 두 집단의 비율이 각각 p_1, p_2 라면 두 비율의 차이는 $p_1 - p_2$ 이며 이는 우리가 평상 적으로 사용하는 비율의 비교 측도이다.
예를 들어 대통령 후보들의 지지율과 차이는 많은 언론에서 사용하고 있으며 기초 통계학에서 두 모집단의 비교를 위한 가설 검정에서도 비율의 차이를 이용하였다.
위의 코로나 치료제의 효과를 비교하는 실험에서 치료집단과 위약집단의 사망률 차이를 측도로 사용하면 어떨까?

2.3.3. 기여위험과 백신효과

기여위험(attributable proportion, attributable risk percent, AR)은 두 그룹의 위험에 대한 비교를 위한 다른 측도이다. 기여위험은 특정한 성격을 가진 집단(exposed group)이 위험에 처한 전체 집단에서 차지하는 비율을 백분율로 나타낸다.

$$AR = \frac{(n_{11}/n_{1+}) - (n_{21}/n_{2+})}{n_{11}/n_{1+}} \times 100 \tag{2.2}$$

예를 들어 비흡연자(unexposed group)와 흡연자(exposed group)의 폐암에 대한 위험을 비교하는 경우를 생각해 보자.비흡연자의 폐암으로 인한 사망률이 연간 1000명 당 0.07명이고 흡연자는 1000명당 0.57명이라고 하면

일단 상대위험은 약 8배이다.

$$RR = 0.57/0.07 = 8.1428$$

두 집단의 비교를 기여위험으로 나타내면 다음과 같다.

$$AR = \frac{0.57 - 0.07}{0.57} \times (100) = 87.7\% \approx 88\%$$

만약 흡연이 폐암을 일으키는 원인이고 두 집단의 다른 요인이 유사하다고 가정하면, 기여위험이 약 88% 라는 것은 모든 폐암 환자(위험에 처한 전체 집단)의 88% 가 흡연에 의한 것이라고 해석할 수 있다.

최근에 코로나 19에 대한 백신과 치료제의 임상실험에서 효과를 발표하는 경우 위에서 언급한 상대위험을 사용하지 않고 **백신효과(Vaccine efficacy, vaccine effectiveness; VE)** 라는 백분율을 사용한다. 백신효과는 기본적으로 기여위험과 동일한 측도이다.

예를 들어 위의 예제에서 치료제의 효과를 백신효과(VE)로 계산하면 다음과 같다.

$$VE = \left[\frac{10/1222 - 5/2360}{10/1222}\right] \times 100 = 74.1101\%$$

백신효과가 74% 란 의미는 치료제를 사용하면 사용하지 않는 경우보다 사망을 74% 줄일 수 있다고 해석할 수 있다.

간단한 예로서 코로나19로 인한 치명율(사망자/확진자)을 비교한다고 가정하자. 백신을 맞은 그룹의 치명율이 1%이고 백신을 맞지 않는 그룹의 치명율이 2% 백신효과는 50%이다.

2.3.4. 오즈비

오드(odd)는 가능성을 나타내는 측도로서 전통적으로 도박에서 유래된 측도이다.

우리가 주사위를 던져서 1과 2가나오면 성공, 다른 숫자가 나오면 실패라고 하는 경우 성공의 확률은 $2/6 = 0.3333$ 으로 계산한다. 확률을 계산하는 경우는 분모에 전체 사건의 수를 사용한다.

위의 주사위 예제로 오드를 계산하면 $2/4 = 0.5$ 가 된다. 즉, 오드는 분모에 성공을 제외한 실패의 사건을 수를 사용한다. 만약 오드가 1이면 무슨 의미인가? 오드가 1이면 성공하는 사건의 수가 실패하는 사건의 수가 동일하다는 의미이다. 게임에서 이길 확률이 1/2 이면 공정한 게임이며 이 경우 오드는 1 이다.

전통적으로 오드는 확률의 개념이 나오기 전에 가능성의 측도로 오랫동안 사용되어 왔으며 도박에서 상대방이 1번 이길 때 내가 이기는 평균적인 횟수를 의미한다.

$$odd = \frac{\text{number of events for success}}{\text{number of events for failure}}$$

예를 들어 위의 코로나 치료제 실험에서 성공을 사망할 사건이라고 하면 위약군의 오드는 $n_{11}/n_{12} = 10/1212$ 이고 치료군의 오드는 $n_{21}/n_{22} = 5/2355$ 이다.

두 집단을 비교하는 측도 중 하나는 **오즈비(odds ratio; OR)**가 있다. 오즈비는 두 그룹의 오드들의 비율로 정의된다. 오즈비가 1이면 두 그룹에서 성공 사건의 가능성이 같다는 것이다.

$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

코로나 치료제 실험에서의 오즈비는 $(10/1212)/(5/2355) = 3.8861$ 이다.

오즈비는 상대위험이나 기여위험에 비하여 의미 있는 해석이 어렵다. 오즈비가 1이면 두 집단이 성공의 가능성이 같다(또는 두 요인의 연관성이 없다)는 것으로 해석이 쉽다. 하지만 예를 들어 오즈비가 1 보다 큰 경우(또는 작은 경우) 두 집단의 차이를 의미 있게 해석하는 것이 어렵다.

오즈비는 향후 학습할 통계적 가설검정에서 중요한 모수(parameter)로 사용되며 특히 실험의 방법이 사례-대조 연구와 같은 특별한 방법을 사용하는 경우 오즈비가 중요한 역할을 하게 된다.

예를 들어 다음과 같은 분할표에서 비율의 차이, 상대위험, 오즈비를 구하여 비교해 보자.

표 2.3.: 2×2 분할표 예제

처리 /결과	성공 ($Y = 1$)	실패 ($Y = 0$)	합계
0 ($X = 0$)	6	4	10
1 ($X = 1$)	4	6	10
합계	10	10	20

비율의 차이(DP)은 다음과 같이 계산된다.

$$DP(0/1) = 6/10 - 4/10 = 0.2$$

상대위험은 다음과 같이 계산된다.

$$RR(0/1) = \frac{6/10}{4/10} = \frac{6}{4} = 1.5$$

오즈비는 다음과 같이 계산된다.

$$OR(0/1) = \frac{6/4}{4/6} = \frac{(6)(6)}{(4)(4)} = 2.25$$

2.4. 신뢰구간

상대위험과 오즈비는 분할표에서 연관성을 나타내는 하나의 측도, 즉 점추정량(point estimation) 이다. 하나의 숫자로 표현되는 점추정은 표본으로 부터 발생한 불확실성을 반영하지 못한다. 따라서 점추정량을 보완하기 위하여 신뢰구간(confidence interval)을 제시할 수 있다.

상대위험과 오즈비는 표본비율 또는 셀 도수의 함수로 나타난다. 하지만 함수의 형태가 비율로서 비선형이기 때문에 상대위험과 오즈비의 근사적인 표준오차(standard error)는 쉽게 구할 수 없다.

다항분포를 가정하고 로그 오즈비의 점근적 분산을 다음과 같이 유도할 수 있다.

$$v_1 = V(\log OR) \approx \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

따라서 로그 오즈비의 $100(1-\alpha)$ % 근사 신뢰구간을 다음과 같이 구할 수 있다.

$$\log OR \pm z_{\alpha/2}\sqrt{v_1}$$

위의 신뢰구간을 오즈비로 역변환하면 오즈비 OR 의 $100(1-\alpha)$ % 근사 신뢰구간을 다음과 같다.

$$(OR \times \exp[-z_{\alpha/2}\sqrt{v_1}],\ \ OR \times \exp[z_{\alpha/2}\sqrt{v_1}]) \tag{2.3}$$

상대위험(RR)의 신뢰구간도 오즈비의 신뢰구간을 유도하는 방법과 유사하게 델타 방법을 사용하며 다음과 같이 구할 수 있다.

$$(RR \times \exp[-z_{\alpha/2}\sqrt{v_2}],\ \ OR \times \exp[z_{\alpha/2}\sqrt{v_2}]) \tag{2.4}$$

위의 식 식 2.4 에서 v_2 는 다음과 같이 계산한다.

$$v_2 = V(\log RR) \approx \frac{1-n_{11}/n_{1+}}{n_{11}} + \frac{1-n_{21}/n_{2+}}{n_{21}}$$

2.4.1. 예제: 아스피린 임상실험

소량의 아스피린 복용이 심장병으로 인한 위험을 줄이는데 효과가 있는지 알아보고자 임상실험을 실시하였다. 22,701명의 남성을 **임의화(randomization)** 을 통해서 두 그룹으로 나눈 후, 한 그룹은 매일 일정량의 아스피린을 복용시키고 다른 그룹은 위약(palcebo)를 복용하게 한 후 약 5년간 심근경색이 일어나는지 알아보았다. 임상실험의 결과는 아래 표와 같다.

표 2.4.: 아스피린 임상실험 결과

	심근경색 발생	심근경색 없음	합
아스피린	139	10,898	11,037
위약	239	10,795	11,034

위약 집단과 아스피린 집단의 상대위험은 다음과 같다.

$$RR = \frac{139/11037}{239/11034} = 0.581$$

상대위험을 보면 1보다 작으므로 아스피린을 복용한 집단이 위약 집단에 비해서 심근 경색이 일어날 위험이 적어진 다는 것을 알 수 있다.

상대위험의 95% 근사 신뢰구간은 다음과 같이 계산한다.

먼저 다음 v_2 를 계산하면

$$v_2 = \frac{1 - n_{11}/n_{1+}}{n_{11}} + \frac{1 - n_{21}/n_{2+}}{n_{21}} = \frac{1 - 139/11037}{139} + \frac{1 - 239/11034}{239} = 0.011$$

상대위험의 신뢰구간은 다음과 같다.

$$(0.581 \times \exp[-1.96\sqrt{0.011}], 0.581 \times \exp[1.96\sqrt{0.011}]) = (0.473, 0.715)$$

위의 신뢰구간은 1을 포함하지 않으므로 상대위험이 1 과 유의한 차이가 있다고 할 수 있다. 결론적으로 아스피린의 복용은 심근경색의 발생을 감소시킨다고 할 수 있다.

이제 `epiR` 패키지를 사용하여 위에서 분석한 내용을 다시 구해보자.

먼저 위의 임상실험 자료를 `R` 의 `matrix` 형태로 저장한다.

```
ex1dat <- matrix( c(139, 10898, 239, 10795), 2, 2, byrow=TRUE)
ex1dat
```

```
     [,1]  [,2]
[1,]  139 10898
[2,]  239 10795
```

이제 함수 `epi.2by2`를 이용하여 상대위험과 상대구간을 구해보자. 임의화를 사용한 임상실험 자료인 경우 `method = "cross.sectional"` 으로 지정한다. 관심이 있는 사건(심근경색, outcome)의 도수가 첫 번째 열(column)에 있으니 `outcome = "as.columns"`이라고 지정한다.

아래 결과에 `Prevalence ratio`라고 나오는 것이 상대위험이다.

```
epi.2by2(dat = ex1dat, method = "cross.sectional", conf.level = 0.95, units = 100,
   interpret = FALSE, outcome = "as.columns")
```

```
             Outcome +    Outcome -      Total               Prev risk *
Exposed +          139        10898      11037       1.26 (1.06 to 1.49)
Exposed -          239        10795      11034       2.17 (1.90 to 2.46)
Total              378        21693      22071       1.71 (1.55 to 1.89)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Prev risk ratio                                 0.58 (0.47, 0.72)
```

```
Prev odds ratio                                    0.58 (0.47, 0.71)
Attrib prev in the exposed *                       -0.91 (-1.25, -0.56)
Attrib fraction in the exposed (%)                -71.99 (-111.63, -39.78)
Attrib prev in the population *                    -0.45 (-0.77, -0.13)
Attrib fraction in the population (%)             -26.47 (-36.51, -17.18)
------------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 26.944 Pr>chi2 = <0.001
Fisher exact test that OR = 1: Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

2.5. 사례-대조 연구

심장발작을 일으킨 환자와 그렇지 않은 사람들을 각각 214명씩 조사하여 과거에 약물남용을 한 경력이 있는지 조사한 사례-대조 연구의 자료이다.

표 2.5.: 약물 남용 사례-대조 연구 결과

	심장 발작 발생	심장발작 없음
약물남용 유	73	18
약물남용 무	141	196
합	214	214

이 연구의 목표는 약물남용과 심장발작의 연관성이 있는지를 알아보는 것이다. 이제 다음과 같은 사건들을 정의해 보자.

- $H+$: 심장발작이 발생했다.
- $H-$: 심장발작이 발생하지 않았다.
- $D+$: 약물남용을 했다.
- $D-$: 약물남용을 하지 않았다.

위에서 정의된 사건들을 고려할 때 사례-대조 연구의 자료에서 다음과 같은 조건부 확률에 대한 추정값을 구할 수 있다.

$$P(\text{약물남용을 했다}|\text{심장발작이 발생했다}) = P(D+|H+) = \frac{73}{214}$$
$$P(\text{약물남용을 하지 않았다}|\text{심장발작이 발생했다}) = P(D-|H+) = 1 - P(D+|H+) = \frac{141}{214}$$

$$P(\text{약물남용을 했다}|\text{심장발작이 발생하지 않았다}) = P(D+|H-) = \frac{18}{214}$$
$$P(\text{약물남용을 하지 않았다}|\text{심장발작이 발생하지 않았다}) = P(D-|H-) = 1 - P(D+|H-) = \frac{196}{214}$$

2.5.1. 사례대조 연구의 목표와 가설

연구에서 비교하고 싶은 비율은 위에서 추정한 확률이 아니고 조건과 결과가 바뀐 다음과 같은 조건부 확률이다.

$$P(\text{심장발작이 발생했다}|\text{약물남용을 했다}) = P(H+|D+)$$
$$P(\text{심장발작이 발생했다}|\text{약물남용을 하지 않았다}) = P(H+|D-)$$

즉 연구의 목표는 다음과 같은 가설을 검정하는 것이다.

$$H_0 : P(H+|D+) = P(H+|D-) \quad \text{vs} \quad H_1 : P(H+|D+) \neq P(H+|D-) \tag{2.5}$$

전체 모집단을 약물남용을 한 사람들과 하지 않은 사람들로 두 집단으로 나누었을 때 두 집단에 대한 심장발작의 확률이 같은지 다른지 비교하고 싶은 것이다.

위의 식에서 보듯이 추정하고 싶은 확률인 $P(H+|D+)$와 $P(H+|D-)$를 추정하려면 전체 모집단에 대한 심장발작 발병률 $P(H+)$와 약물남용의 비율 $P(D+)$를 알아야 한다. 즉

$$\begin{aligned} P(H+|D+) &= \frac{P(H+ \cap D+)}{P(D+)} \\ &= \frac{P(D+|H+)P(H+)}{P(D+)} \\ &\approx (73/214)\frac{P(H+)}{P(D+)} \end{aligned}$$

위의 식은 다음의 조건부 확률 공식을 각 단계마다 적용한 결과이다.

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

사례-대조 연구의 자료만으로는 모집단에 대한 심장발작 발병률 $P(H+)$와 약물남용의 비율 $P(D+)$을 구할 수 없다. 또한 다른 외부의 자료가 있다 하더라도 약물남용의 비율을 정확하게 추정하는 것은 매우 어렵다.

2.5.2. 오즈비의 비교

이러한 문제는 두 집단의 비율의 차이나 상대위험을 비교하지 않고 오즈비를 구하여 비교하면 심장발작 발병률와 약물남용의 비율을 추정하지 않고 사례-대조 연구의 자료만으로 추론이 가능하다.

다음의 가설은 두 비율의 비교를 오즈비로 표현한 것이다.

$$H_0 : \frac{P(H+|D+)/P(H-|D+)}{P(H+|D-)/P(H-|D-)} = 1 \quad \text{vs} \quad H_1 : \frac{P(H+|D+)/P(H-|D+)}{P(H+|D-)/P(H-|D-)} \neq 1 \tag{2.6}$$

위의 가설 식 2.6 는 단순한 비율을 비교하는 가설 식 2.5 과 동일한 가설이다.

가설 식 2.6 에서 나타는 오즈비는 심장발작 발병률와 약물남용의 비율을 이용하지 않고 사례-대조 연구에서 추정할 수 있는 조건부 확률만으로 추정할 수 있다.

$$\begin{aligned}
\frac{P(H+|D+)/P(H-|D+)}{P(H+|D-)/P(H-|D-)} &= \frac{[P(H+|D+)P(D+)]/[P(H-|D+)P(D+)]}{[P(H+|D-)P(D-)]/[P(H-|D-)P(D-)]} \\
&= \frac{P(H+\cap D+)/P(H-\cap D+)}{P(H+\cap D-)/P(H-\cap D-)} \\
&= \frac{[P(D+|H+)P(H+)]/[P(D+|H-)P(H-)]}{[P(D-|H+)P(H+)]/[P(D-|H-)P(H-)]} \\
&= \frac{P(D+|H+)/P(D+|H-)}{P(D-|H+)/P(D-|H-)} \\
&= \frac{(73/214)/(142/214)}{(18/214)/(196/214)} \\
&= \frac{(73)(196)}{(141)(18)} \\
&= 5.64
\end{aligned}$$

결론적으로 사례-대조 연구에서는 연구의 목표에 대한 가설 검정을 비율의 차이나 상대위험으로 표현하여 수행할 수 없다. 하지만 오즈비를 검정하는 것으로 가설을 세우면 자료에서 쉽게 유도할 수 있는 오즈비로 가설 검정을 쉽게 수행할 수 있다.

2.5.3. 예제: 약물남용 사례-대조 연구

심작발작을 일으킨 환자와 그렇지 않은 사람들을 각각 214명씩 조사하여 과거에 약물남용을 한 경력이 있느지 조사한 **사례-대조 연구(case-control study)** 의 결과가 표 표 2.5 에 있다.

사례-대조 연구는 사례(case)가 발견되면, 즉 위의 연구와 같이 심장발작이 일어난 환자가 발생하면 그 환자와 유사한 나이와 성별 등을 가진 일반사람을 찾아 매칭하여 환자와 일반인의 과거 경력을 조사하는 후향적인 연구(restrspective study)이다. 반대로 앞의 예제에서 본 임의화를 이용한 임상실험은 전향적 연구(prospective study)이다.

이러한 사례-대조 연구에서는 상대위험을 이용하여 연관성을 알아낼 수 없다. 하지만 사례-대조 연구에서 상대위험 대신 오즈비를 이용하여 연관성을 추론할 수 있다.

위의 심장발작에 대한 사례-대조 연구의 결과에서 오즈비와 그 신뢰구간을 구해보자.

먼저 오즈비는 다음과 같다.

$$OR = \frac{(73)(196)}{(18)(141)} = 5.64$$

위의 결과는 심장발작이 일어난 집단에서 약물남용을 한 환자들의 오즈가 심장발작이 일어나지 않은 집단에서 약물남용을 한 사람들의 오즈에 비해 5.6배 크다는 것을 알 수 있으며 이는 1보다 상당히 크다.

오즈비의 95% 근사 신뢰구간은 다음과 같이 계산한다.

먼저 다음 v_1 를 계산하면

$$v_1 = V(\log OR) \approx \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} = \frac{1}{73} + \frac{1}{18} + \frac{1}{141} + \frac{1}{196} = 0.08$$

상대위험의 신뢰구간은 다음과 같다.

$$(5.64 \times \exp[-1.96\sqrt{0.08}], 5.64 \times \exp[1.96\sqrt{0.08}]) = (3.222, 9.863)$$

위의 신뢰구간을 보면 1을 포함하지 않으므로 약물남용이 심장발작의 위험을 높인다고 말할 수 있다.

이제 `epiR` 패키지를 사용하여 위에서 분석한 내용을 다시 구해보자.

먼저 위의 사례-대조 연구 자료를 R 의 `matrix` 형태로 저장한다.

```
ex2dat <- matrix( c(73,18,141,196), 2, 2, byrow=TRUE)
ex2dat
```

```
     [,1] [,2]
[1,]   73   18
[2,]  141  196
```

이제 함수 `epi.2by2`를 이용하여 오즈비과 상대구간을 구해보자. 사례-대조 연구의 자료인 경우 `method = "case.control"`으로 저장한다. 사례-대조 연구로 지정하면 상대위험이 출력되지 않는다. 관심이 있는 사건(심장발작, outcome)의 도수가 첫 번째 열(column)에 있으니 `outcome = "as.columns"`이라고 지정한다.

```
epi.2by2(dat = ex2dat, method = "case.control", conf.level = 0.95, units = 100,
   interpret = FALSE, outcome = "as.columns")
```

```
             Outcome +    Outcome -      Total                        Odds
Exposed +           73           18         91        4.06 (2.50 to 7.27)
Exposed -          141          196        337        0.72 (0.57 to 0.89)
Total              214          214        428        1.00 (0.83 to 1.21)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Exposure odds ratio                             5.64 (3.22, 9.86)
Attrib fraction (est) in the exposed (%)       82.19 (68.26, 90.44)
Attrib fraction (est) in the population (%)    28.06 (20.13, 35.21)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 42.218 Pr>chi2 = <0.001
Fisher exact test that OR = 1: Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
```

3. 연관성의 검정

이 절에서는 두 변수의 연관성에 통계적 가설 검정 방법을 살펴보자.

3.1. 필요한 패키지

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
```

3.2. 카이제곱 검정

일단 2개의 이항변수 X 와 Y 를 고려하고 가능한 결과의 조합과 그 확률은 다음과 같은 2×2 분할표로 나타낼 수 있다.

		Y		
		0	1	Total
X	0	p_{11}	p_{12}	p_{1+}
	1	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	1

그림 3.1.: 2 x 2 분할표

일반적으로 2×2 분할표에서 다음과 같은 두 가지 가설이 가능하다.

- 동질성 검정(homogeneuty test)

  변수 X 가 단순하게 독립 집단을 나누는 변수인 경우 (예를 들어 실험약 집단과 위약 집단) 두 그룹 간에 이항변수 Y 의 성공확률이 같은지 검정하는 경우이다. 실험약 집단과 위약 집단에서 심장병이 발병할 확률이 같은지 검정을 수행할 때 귀무가설은 다음과 같다.

$$H_0 : p_{1j} = p_{2j} = p_j$$

- 독립성 검정(independent test)

  변수 X 와 Y 가 모두 확률변수인 경우 두 변수가 독립인지 검정하는 경우이다. 예를 들어 흡연(X)과 심근경색(Y)의 관계를 연구하는 경우 두 사건이 모두 확률적인 사건이라고 보고 다음과 같이 독립에 대한 가설을 고려한다.

$$H_0 : p_{ij} = p_{i+}p_{+j}$$

다음과 같이 n 개의 관측값으로 구성된 2×2 분할표에서 **동질성과 독립성 가설을 검정하는 방법은 동일하며** 따라서 굳이 두 가지 가설을 엄격하게 구별할 이유는 없다. 만약 귀무가설이 기각되면 두 변수의 연과성은 유의하다고 결론을 내린다.

		Y		
		0	1	Total
X	0	n_{11}	n_{12}	n_{1+}
	1	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

그림 3.2.: 2 x 2 분할표: 관측 도수

동질성과 독립성에 대한 검정은 다음과 같은 **카이제곱 통계량**을 사용한다.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_i j} \tag{3.1}$$

위의 카이제곱 통계량에서 $O_{ij} = n_{ij}$ 는 각 셀의 관측도수이며 E_{ij}는 귀무가설 하에서의 셀 도수의 예측값이다.
동질성 검정을 고려할 때 만약 귀무가설이 참이라면 확률 $p_{1j} = p_{2j} = p_j$ 는 다음과 같이 추정할 수 있다.

$$\hat{p}_j = \frac{n_{+j}}{n}$$

따라서 셀 (i, j) 에 대한 기데 돗수 E_{ij} 는 다음과 같이 계산된다.

$$E_{ij} = n_{i+}\hat{p}_j = \frac{n_{i+}n_{+j}}{n} \tag{3.2}$$

귀무가설 하에서 표본의 크기가 충분히 크면 식 식 3.1 의 카이제곱 검정통계량 χ^2 는 자유도가 1인 카이제곱 분포를 따른다. 그러므로 이 사실을 이용하여 p-값을 계산하거나 기각역을 구하여 검정한다.

일반적인 $I \times J$ 분할표도 동일한 방법으로 가설검정을 할 수 있다. 카이제곱 통계량을 구하는 방법은 2×2 분할표와 유사하다. 다만 귀무가설이 참인 경우 검정통계량은 자유도가 $(I-1)(J-1)$ 인 카이제곱 분포를 따른다.

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_i j}$$

이제 실제 분할표에서 카이제곱 검정을 수행해 보자. 아스피린 임상실험 결과가 주어진 표 표 2.4 에서 아스피린의 효과사 없는 경우, 즉 귀무가설이 참인 경우 다음과 같이 심근경색의 유무에 대한 예측 확률을 구할 수 있다.

$$\hat{p}_1 = \frac{n_{+1}}{n} = \frac{139 + 239}{22071} = 0.0171$$
$$\hat{p}_2 = \frac{n_{+2}}{n} = \frac{10898 + 10795}{22071} = 0.9829$$

이제 각 셀의 기대도수를 식 식 3.2 에 의하여 계산할 수 있다. 예를 들어 E_{11} 은 다음과 같이 계산된다.

$$E_{11} = \frac{n_{1+}n_{+1}}{n} = n_{1+}\hat{p}_1 = (11037)(0.0171) = 189.03$$

각 셀에 대한 기대도수 E_{ij} 를 구하고 식 식 3.1 의 카이제곱 통계량을 구하면 다음과 같다.

$$\begin{aligned}\chi^2 &= \frac{(139 - 189.03)^2}{189.03} + \frac{(10898 - 10848.00)^2}{10848.00} \\ &\quad + \frac{(239 - 188.97)^2}{188.97} + \frac{(10795 - 10845.03)^2}{10845.03} \\ &= 26.94\end{aligned}$$

자유도가 1인 카이제곱 분포의 상위 5% 백분위수 3.84 이다. 위에서 구한 카이제곱 통계량의 값이 26.94 로서 3.84 보다 크므로 귀무가설을 기각한다. 즉 아스프린과 위약을 복용한 두 그룹 사이에는 심근경색이 일어날 비율에 유의한 차이가 있다.

R 에서도 카이제곱 검정을 쉽게 수행할 수 있다. 앞에서 표 표 2.4 의 자료를 행렬의 형태로 저장하였는데 함수 `chisq.test()` 를 사용하면 결과를 쉽게 구할 수 있다.

```
ex1dat <- matrix( c(139, 10898, 239, 10795), 2, 2, byrow=TRUE)
ex1dat
```

```
     [,1]  [,2]
[1,]  139 10898
[2,]  239 10795
```

```
chisq.test(ex1dat)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  ex1dat
X-squared = 26.408, df = 1, p-value = 2.764e-07
```

분할표에서의 기대도수 E_{ij} 는 다음과 같이 얻을 수 있다.

```
chisq.test(ex1dat)$expected
```

```
         [,1]     [,2]
[1,] 189.0257 10847.97
[2,] 188.9743 10845.03
```

3.3. 코크란-맨텔-헨젤 검정

임상실험이나 의학연구는 여러 나라 또는 여러 병원들에서 진행되는 경우가 있다. 이러한 경우 국가나 병원의 고유한 특성에 따라서 실험의 결과가 다르게 나타날 수 있다. 이렇게 그룹에 의한 효과를 그룹 효과 또는 층(strata)에 의한 효과라고 한다. 예를 들어 진통제에 대한 효과는 그 나라의 문화나 관습에 따라서 효과의 차이가 나타날 수 있다. 또한 여러 개의 변원에서 연구가 동시에 진행된다면 병원의 규모, 위치, 환자들의 특성에 따라서 치료 효과의 차이가 나타날 수 있다.

이렇게 그룹에 따른 차이가 예상되는 경우 그룹의 효과를 제어하면서 처리 효과의 차이를 검정하는 방법이 필요하다. 이렇게 여러 개의 층으로 구성된 독립집단에서 얻은 자료에서 층에 의한 횩과를 통제하면서 동질성 또는 독립성 검정을 수행하는 방법을 **코크란-맨텔-헨젤 검정 (Cochran-Mantel-Haenzel test)**라고 한다.

아래와 같이 K 개의 독립집단(또는 층)에서 각각 얻은 K 개의 2×2 분할표가 있다고 하자.

		Y		
		0	1	Total
X	0	n_{k11}	n_{k12}	n_{k1+}
	1	n_{k21}	n_{k22}	n_{k2+}
	Total	n_{k+1}	n_{k+2}	n_k

그림 3.3.: K 개의 2 x 2 분할표

K 개의 독립집단이 있고 성공의 확률이 p_1, 실패의 학률이 p_2 라고 한다면 처리의 효과를 전체적으로 비교하는 가설은 다음과 같다.

$$H_0 : p_1 = p_2 \quad \text{sv.} \quad H_1 : p_1 \neq p_2$$

이제 귀무가설의 가정 하에서 각 분할표에서 n_{k11} 에 대한 기대도수 μ_{k11} 와 그 분산 v_{k11} 을 다음과 같이 계산한다.

$$\mu_{k11} = E(n_{k11}|H_0) = \frac{n_{k1+}n_{k+1}}{n_k}$$

$$v_{k11} = V(n_{k11}|H_0) = \frac{n_{k1+}n_{k2+}n_{k+1}n_{k+2}}{n_k^2(n_k - 1)}$$

이제 가설검정을 위한 통계량 Q_{CMH} 은 다음과 같다.

$$Q_{CMH} = \frac{\left[\sum_{k=1}^{K}(n_{k11} - \mu_{k11})\right]^2}{\sum_{k=1}^{K} v_{k11}} \tag{3.3}$$

귀무가설이 참인 경우 검정통계량 Q_{CMH} 은 자유도가 1 인 카이제곱 분포를 따른다.

이제 Agresti (2012) 의 6.3절에 있는 **다기관 임상시험(multi-center clinical trial)** 의 예제를 살펴보자. 아래 표는 모두 8개의 독립적인 병원에서 감염 치료제에 대한 효과에 대한 실험을 실시하여 얻은 자료이다.

TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	var(n_{11k})
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

그림 3.4.: 8개 병원의 임상실험 결과

마지막 병원을 제외한 7개의 병원에서 치료제의 효과가 긍정적으로 나타났다. 여기서 주목할 점은 병원에 따라서 연관성의 강도가 매우 다르게 나타날 수 있다는 것이다.

이제 각 병원을 층(strata)로 고려하고 병원의 효과를 제어하면서 식 식 3.3 의 검정 통계량 Q_{CMH} 를 이용하여 치료제의 효과가 있는지 검정해보자. 검정은 아래와 같이 R 프로그램을 이용한다. 함수 `mantelhaen.test()` 는 코크란-맨텔-헨젤 검정을 수행하는 함수이다.

```
beitler <- c(11,10,25,27,16,22,4,10,14,7,5,12,2,1,14,16,6,0,11,12,1,0,10,10,1,1,4,8,4,6,2,1)
beitler <- array(beitler, dim=c(2,2,8))
beitler
```

```
, , 1
```

```
     [,1] [,2]
[1,]   11   25
[2,]   10   27

, , 2

     [,1] [,2]
[1,]   16    4
[2,]   22   10

, , 3

     [,1] [,2]
[1,]   14    5
[2,]    7   12

, , 4

     [,1] [,2]
[1,]    2   14
[2,]    1   16

, , 5

     [,1] [,2]
[1,]    6   11
[2,]    0   12

, , 6

     [,1] [,2]
[1,]    1   10
[2,]    0   10

, , 7

     [,1] [,2]
[1,]    1    4
[2,]    1    8

, , 8

     [,1] [,2]
[1,]    4    2
[2,]    6    1
```

```
mantelhaen.test(beitler, correct=FALSE)
```

```
    Mantel-Haenszel chi-squared test without continuity correction

data:  beitler
Mantel-Haenszel X-squared = 6.3841, df = 1, p-value = 0.01151
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.177590 3.869174
sample estimates:
common odds ratio
         2.134549
```

검정 통계량 Q_{CMH} 의 값이 6.3841 이고 p-값은 0.0115 이므로 귀무가설을 기각한다.

3.4. 맥나마 검정

연속형 변수에서 짝지은 자료를 비교할 때 사용하는 방법이 대응 t-검정(paired t-test) 또는 짝표본 t-검정이다. 예를 들어 천식환자가 A약을 먹고 폐활량을 측정하고 일정 기간이 지나서 같은 환자가 B약을 먹고 폐활량을 측정하면 두 관측값은 독립이 아니다. 따라서 이러한 경우 독립 t-검정이 아닌 대응 t-검정을 사용한다.

이제 이산형 변수가 짝으로 나타나는 경우를 생각해보자. 예를 들어 눈병 치료에 사용되는 A약과 B약의 효과를비교하기 위하여 각각의 약을 환자의 오른쪽 눈과 왼쪽 눈에 처치를 하고 치료의 여부를 관측하였다고 하자.

		Right eye		
		cured	not cured	Total
Left eye	cured	n_{11}	n_{12}	n_{1+}
	not cured	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

그림 3.5.: 짝표본 실험에 의한 2 x 2 분할표

위의 표에서 n_{11} 은 A약과 B약의 효과가 모두 나타난 환자의 도수이다. n_{12} 은 A약은 효과가 있고 B약은 효과가 없는 환자의 도수이다. 이러한 자료는 앞에서 배운 카이제곱 검정을 적용할 수 없다.

이제 일반적으로 짝표본에서 나온 자료가 다음 표와 같이 얻어졌다고 가정하자.

이제 조건 1 에서 성공의 확률을 p_1 이라고 하고 조건 2에서 성공의 확률을 p_2 라고 하면 짝표본에서 얻어진 분할표 그림 3.6 에서 관심있는 가설은 다음과 같다.

$$H_0 : p_1 = p_2 \quad \text{sv.} \quad H_1 : p_1 \neq p_2$$

분할표 그림 3.6 에서 p_1 과 p_2 의 추정량은 다음과 같다.

		조건 2		
		예	아니오	Total
조건 1	예	n_{11}	n_{12}	n_{1+}
	아니오	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

그림 3.6.: 짝표본 실험에 의한 2 x 2 분할표

$$\hat{p}_1 = \frac{n_{1+}}{n}, \quad \hat{p}_2 = \frac{n_{+1}}{n}$$

p_1 과 p_2 의 추정량의 차이는 두 조건에 따른 결과가 일치하지 않는 도수 n_{12} 와 n_{21} 의 차이에 의존한다.

$$\hat{p}_1 - \hat{p}_2 = \frac{n_{1+}}{n} - \frac{n_{+1}}{n} = \frac{n_{11}+n_{12}}{n} - \frac{n_{11}+n_{21}}{n} = \frac{n_{12}-n_{21}}{n}$$

맥나마 검정(McNemar Test)는 도수 n_{12} 와 n_{21} 에 의거하여 두 확률이 같은지 검정하는 방법을 제시하였다. 맥나마 검정을 위한 통계량은 다음과 같다.

$$Q_M = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}} \tag{3.4}$$

맥나마 검정 통계량 Q_M 은 귀무가설 하에서 근사적으로 자유도가 1인 카이제곱 분포를 따른다.

다음은 1600명 영국 시민들의 수상에 대한 지지 여부를 두 개의 연속된 여론 조사에서 수집한 자료이다 (Agresti 2012 의 10 장 참조). 이제 두 시점에서 수상에 대한 지지율이 같은지 아닌지 R 을 이용하여 맥나마 검정을 해보자. 맥나마 검정은 함수 `mcnemar.test()`를 사용하여 수행할 수 있다.

TABLE 10.1 Rating of Performance of Prime Minister

First Survey	Second Survey: Approve	Second Survey: Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

그림 3.7.: 영국시민의 수상에 대한 지지도 조사 자료

```
ex3dat <- matrix(c(794,150,86,570),byrow=T,ncol=2)
ex3dat
```

```
     [,1] [,2]
[1,]  794  150
```

```
[2,]   86  570
```

```
mcnemar.test(ex3dat ,correct=F)
```

```
    McNemar's Chi-squared test

data:  ex3dat
McNemar's chi-squared = 17.356, df = 1, p-value = 3.099e-05
```

검정의 p-값이 매우 작으므로 귀무가설을 기각한다. 두 시점에서 수상에 대한 지지율이 하락했다고 할 수 있다. 참고로 첫 번째 조사에서의 지지율의 추정치는 $\hat{p}_1 = 944/1600 = 0.59$ 이고 두 번째 조사에서의 지지율의 추정치는 $\hat{p}_2 = 880/1600 = 0.55$ 이다. 또한 의견을 바꾸지 않은 사람의 비율은 $(794 + 570)/1600 = 0.8225$로 대부분의 시민들이 지지 의견을 바꾸지 않았다.

4. 진단의 평가

의학에서 진단은 환자의 상태나 질병의 징후를 판단하는 일이다. 진단을 수행하기 위해서 의사는 전통적인 진단법도 사용하지만 다양한 계측 기계를 이용하는 진단 기법도 사용한다. 최근에는 첨단 분석 장비를 이용하여 다양한 질병에 대한 진단을 정확하고 쉽게 할 수 있다. 특히 최근 코로나 시대에 들면서 일반인도 여러 가지 이유로 진단 검사를 받는 경우가 자주 일어난다.

진단 기법을 사용하여 감염 여부 등을 판단하는 경우 언제나 오류가 발생한다. 최근에는 첨단 기술 덕분으로 이러한 오류율이 많이 줄어 들었지만 오류가 없는 진단 기법은 개발하기 힘들다.

코로나 검사를 받고 음성 판정을 받아도 실제 양성인 경우가 나타나며, 반대로 양성 판정을 받아도 음성이 경우가 나타난다. 이렇게 **진단에서 발생하는 오류는 두 가지 종류** 가 있다.

연구자들이 진단 기법을 개발할 때 오류의 가능성이 작아지도록 노력하지만, 불행하게도 두 가지 오류의 확률을 모두 0 으로 만들 수 없다.

극단적인 예를 들어보자. 코로나 바이러스 감염의 유무를 판단하는 진단 기법 A는 검사를 받는 사람을 모두 양성이라고 판단한다고 하자. 이 경우 양성인 사람이 음성으로 잘못 판단되는 오류의 확률은 0이다. 반대로 진단 기법 B는 검사를 받는 사람을 모두 음성이라고 판단한다면 음성인 사람이 양성으로 잘못 판단되는 오류의 확률은 0이다. 여기서 진단 기법 A 와 B 는 모두 쓸모없는 검사라는 것을 우리는 잘 알고 있다. 양성인 사람과 음성인 사람을 잘 구별할 수 있는 진단 기법이 좋은 방법이다.

이제 우리는 진단 기법을 평가할 때 사용되는 확률의 측도에 대하여 알아보자.

> **노트**
>
> 일반적으로 양성(positive)는 바이러스에 감염되었거나 질병이 있다는 사건을 말한다. 음성(negative)는 양성의 반대 사건이다. 하지만 양성과 음성의 의미가 바뀌는 경우도 종종 있다.

4.1. 민감도와 특이도

진단 기법을 평가하는 경우 다음과 같은 두 질문에 대해서 생각해 보아야 한다.

- 양성인 사람을 얼마나 잘 양성으로 판단하는가?
- 음성인 사람을 얼마나 잘 음성으로 판단하는가?

양성인 사람을 얼마나 잘 양성으로 판단하는지에 대한 평가 기준이 **민감도(sensitivity)** 이고 음성인 사람을 얼마나 잘 음성으로 판단하는지에 대한 평가 기준이 **특이도(specificity)** 이다. 민감도와 특이도의 정도는 확률로서 나타낼 수 있다.

진단 기법에 대한 실험 연구를 수행하면 그 결과는 2×2 분할표로 다음과 같이 요약할 수 있다. 일반적으로 진단 기법의 효과를 측정하는 실험은 대상자에 대한 질병의 유무를 알고 시작한다.

표 4.1.: 진단 기법의 실험 결과

진단(T) / 질병(D)	양성 ($D+$)	음성 ($D-$)
양성 ($T+$)	TP	FP
음성 ($T-$)	FN	TN

위의 표에서 각 셀에 해당하는 진단 결과는 다음과 같이 나타낼 수 있다.

- TP : True Positive
- FP : False Positive
- FN : False Nagative
- TN : True Negative

이제 분할표 표 4.1 에서 민감도와 특이도는 다음과 같이 정의된다.

$$\text{Sensitivity(민감도)} = \frac{TP}{TP + FN} \tag{4.1}$$

$$\text{Specificity(특이도)} = \frac{TN}{FP + TN} \tag{4.2}$$

다음은 코로나 바이러스 검사법에 대한 여러 연구에서 나온 민감도와 특이도 결과를 보여 준다 (Butler-Laporte 기타 (2021)).

Figure 3. Primary Meta-analysis Results for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2 in Saliva Samples

Source	Test finding, No. TP	FP	FN	TN	Sensitivity (95% CrI), %	Specificity (95% CrI), %
Akgun et al[14]	30	5	25	38	60.6 (46.3-77.0)	98.8 (95.4-99.9)
Becker et al[15]	11	2	10	62	65.4 (44.3-89.8)	99.0 (96.5-99.9)
Byrne et al[16]	12	0	2	96	82.2 (64.8-95.5)	99.3 (98.0-99.9)
Caulley et al[22]	34	14	22	1869	79.5 (56.8-95.8)	99.3 (98.3-99.9)
Cheuk et al[25]	104	37	18	70	84.5 (72.4-92.0)	99.1 (97.3-99.9)
Iwasaki et al[26]	8	1	1	66	83.8 (65.8-96.1)	99.2 (97.5-99.9)
Hanson et al[23]	75	6	5	268	87.3 (74.1-96.5)	99.2 (97.7-99.9)
Kojima et al[17]	20	6	3	16	84.6 (69.9-95.0)	99.1 (97.2-99.9)
Landry et al[27]	28	2	5	89	81.8 (66.8-93.9)	99.2 (97.6-99.9)
McCormick-Baw et al[29]	47	1	2	105	86.0 (71.1-96.0)	99.2 (97.8-99.9)
Miller et al[21]	33	1	1	56	86.7 (73.1-96.6)	99.2 (97.6-99.9)
Pasomsub et al[3]	16	2	3	179	83.3 (66.1-96.0)	99.2 (97.8-99.9)
Teo et al[18]	139	70	11	116	89.4 (77.1-96.3)	99.1 (97.1-99.9)
Vogels et al[19]	32	3	2	30	86.4 (72.1-96.1)	99.1 (97.2-99.9)
Williams et al[28]	33	1	6	49	79.9 (65.0-91.6)	99.1 (97.5-99.9)
Yokata et al[20]	42	6	4	1872	85.9 (71.0-96.6)	99.2 (98.2-99.8)
Pooled	**664**	**157**	**120**	**4981**	**83.2 (74.7-91.4)**	**99.2 (98.2-99.8)**

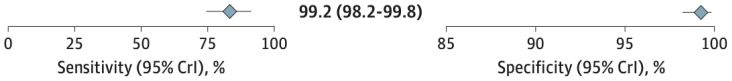


그림 4.1.: 코로나 검사의 민감도와 특이도

예제로서 그림 그림 4.1 에 제시한 종합적인 결과(pooled counts)를 이용하여 민감도와 특이도를 구해보자.

표 4.2.: 코로나 바이러스 검사법의 결과

진단(T) / 질병(D)	양성 ($D+$)	음성 ($D-$)
양성 ($T+$)	664	157
음성 ($T-$)	120	4981

민감도와 특이도는 다음과 같이 구할 수 있다.

$$\text{Sensitivity} = \frac{664}{664+120} = 0.8469$$
$$\text{Specificity} = \frac{4891}{157+4891} = 0.9689$$

위에서 구한 민감도와 특이도는 Butler-Laporte 기타 (2021) 에서 제시한 민감도(83.2%), 특이도(99.2%) 와 유사하지만 약간의 차이가 있다. 그 이유는 Butler-Laporte 기타 (2021) 는 모든 실험 결과를 단순하게 더한 것이 아니라 메타분석(meta analysis)을 사용하여 얻은 결과이기 때문이다. 메타분석은 같은 주제에 대한 여러 개의 독립적인 연구 결과들을 결합하여 결론을 추론하는 연구 방법이다.

4.2. 양성예측도와 음성예측도

앞에서 살펴본 민감도와 특이도를 구하는 실험에서는 실험 대상자가 질병이 있는지 없는지 알고 있다. 하지만 실제 검사는 진단을 받는 사람이 질병이 있는지 모르는 상태에서 진행된다.
따라서 우리가 정말 관심 있는 확률은 **양성으로 진단된 사람이 실제로 양성인지?**에 대한 확률이다.

양성으로 판정되었을 때 실제로 병에 걸렸을 확률을 **양성예측도**($PV+$) (predicted value of positive test, predictive value positive) 라고 부르며 음성으로 판정되었을 때 실제로 병에 걸리지 않았을 확률을 **음성예측도**($PV-$) (predicted value of negative test, predicted value negative) 라고 부른다. 양성예측도와 음성예측도는 조건부 확률로 표현할 수 있다.

$$PV+ = P(D+|T+) \tag{4.3}$$
$$PV- = P(D-|T-) \tag{4.4}$$

이제 앞에서 살펴본 민감도와 특이도도 다음과 같이 조건부 확률로 나타낼 수 있다.

$$\text{Sensitivity} = P(T+|D+) \tag{4.5}$$
$$\text{Specificity} = P(T-|D-) \tag{4.6}$$

이제 실제로 중요한 양성예측도와 음성예측도를 민감도와 특이도를 이용하여 유도해 보자. 두 확률은 사건과 조건이 바뀐 확률이기 때문에 베이즈 정리(Bayes' Theorem)을 이용하여 구할 수 있다.

일단 양성예측도를 구하는 식을 베이즈 정리를 적용하여 유도해 보자.

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

위의 식에서 나타나는 확률 $P(D+)$ 는 모집단에서 질병에 걸린 사람들의 비율을 의미하며 이를 **유병률(prevalence)** 이라고 부른다. 즉 양성예측도를 구하려면 질병의 유병률을 알아야 한다.

다시 식을 정리해 보면 양성예측도에 대한 공식은 다음과 같다.

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} && (4.7)\\ &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + [1-P(T-|D-)][1-P(D+)]} && (4.8)\\ &= \frac{(\text{민감도})(\text{유병률})}{(\text{민감도})(\text{유병률}) + (1-\text{특이도})(1-\text{유병률})} && (4.9) \end{aligned}$$

비슷한 계산 방법으로 음성예측도는 다음과 같이 주어진다.

$$\begin{aligned} P(D-|T-) &= \frac{P(T-|D-)P(D-)}{P(T-|D-)P(D-) + P(T-|D+)P(D+)} && (4.10)\\ &= \frac{P(T-|D-)[1-P(D+)]}{P(T-|D-)[1-P(D+)] + [1-P(T+|D-)]P(D+)} && (4.11)\\ &= \frac{(\text{특이도})(1-\text{유병률})}{(\text{특이도})(1-\text{유병률}) + (1-\text{민감도})(\text{유병률})} && (4.12) \end{aligned}$$

이제 표 표 4.2 의 결과를 이용하여 코로나 검사의 양성예측도와 음성예측도를 구해보자.

코로나 유병률은 나라마다 다르고 추정하기도 힘들다. 따라서 쉽게 현재 까지 누적환자수를 전체 인구로 나눈 단순한 비율을 유병률로 사용해 보자(주의! 우리가 여기서 사용한 비율은 실제 유병률을 계산하는 방법과 다르다). 2021년 현재 누적 환자 수가 274,415 명이고 2020년 기준 총인구는 51,829,136 명이므로 유병률을 $274415/51829136 = 0.0053$ 이라고 하자.

이제 표 표 4.2 의 결과를 이용하면 코로나 검사의 양성예측도와 음성예측도는 다음과 같이 추정할 수 있다.

$$\begin{aligned} P(D+|T+) &= \frac{(\text{민감도})(\text{유병률})}{(\text{민감도})(\text{유병률}) + (1-\text{특이도})(1-\text{유병률})}\\ &= \frac{(0.8469)(0.0053)}{(0.8469)(0.0053) + (1-0.9689)(1-0.0053)}\\ &= 0.1267 \end{aligned}$$

```
(0.8469 )*(0.0053)/((0.8469 )*(0.0053) + (1-0.9689)*(1-0.0053))
```

```
[1] 0.1267108
```

$$P(D-|T-) = \frac{(\text{특이도})(1-\text{유병률})}{(\text{특이도})(1-\text{유병률}) + (1-\text{민감도})(\text{유병률})} \tag{4.13}$$

$$= \frac{(0.9689)(1-0.0053)}{(0.9689)(1-0.0053) + (1-0.8469)(0.0053)} \tag{4.14}$$

$$= 0.9992 \tag{4.15}$$

```
(0.9689)*(1- 0.0053)/((0.9689)*(1-0.0053) + (1-0.8469)*(0.0053))
```

```
[1] 0.9991588
```

사실 코로나 유병률은 정확하게 알 수도 없고 시간에 따라 변할 것이다. 이제 다양한 유병률에 따라서 양성예측도와 음성예측도가 어떻게 변하는지 계산해 보자.

```
calpred <- function(prev, sen, spe){
    pred.pos <- sen*prev/(sen*prev + (1-spe)*(1-prev))
  pred.neg <- spe*(1-prev)/(spe*(1-prev) + (1-sen)*(prev))
  res <- data.frame(sen, spe, prev, pred.pos, pred.neg)
  colnames(res) <- c("Sensitivity", "SPecificity","Prevalnce", "Pred. Post.", "Pred. Nega.")
  res
}

preval.range <- seq(0, 0.02, 0.002)
calpred(preval.range ,0.8469, 0.9689 )
```

```
   Sensitivity SPecificity Prevalnce Pred. Post. Pred. Nega.
1       0.8469      0.9689     0.000  0.00000000   1.0000000
2       0.8469      0.9689     0.002  0.05174816   0.9996834
3       0.8469      0.9689     0.004  0.09858220   0.9993658
4       0.8469      0.9689     0.006  0.14117039   0.9990471
5       0.8469      0.9689     0.008  0.18006506   0.9987273
6       0.8469      0.9689     0.010  0.21572673   0.9984064
7       0.8469      0.9689     0.012  0.24854242   0.9980845
8       0.8469      0.9689     0.014  0.27883973   0.9977614
9       0.8469      0.9689     0.016  0.30689786   0.9974372
10      0.8469      0.9689     0.018  0.33295620   0.9971120
11      0.8469      0.9689     0.020  0.35722119   0.9967856
```

```
calpred(0.0053, 0.8469, 0.9689  )
```

```
  Sensitivity SPecificity Prevalnce Pred. Post. Pred. Nega.
1      0.8469      0.9689    0.0053   0.1267108   0.9991588
```

Part II.

일반화 선형모형

5. 로지스틱 회귀모형

5.1. 필요한 패키지

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
library(alr4)
library(sm)
library(MASS)
library(knitr)
library(kableExtra)
```

5.2. 이항변수: 예제

먼저 R을 이용한 로지스틱 회귀분석을 소개하기 위하여 다음의 예제를 이용하고자 한다.

5.2.1. 챌린져호 O-ring 자료

1986년 미국우주항공국(NASA)이 발사한 우주왕복선 챌린져호(Spache Shuttle Challenger)가 로켓 엔진에 주요부품인 O-rings의 손상으로 인하여 공중에서 폭팔하는 사고가 일어났다. 다음 데이타는 미국우주항공국이 챌린져호를 발사하기 전에 실험을 통하여 얻은자료이다. 디음의 자료는 교과서의 R package `faraway`에서 `orings`에서 볼 수 있다.

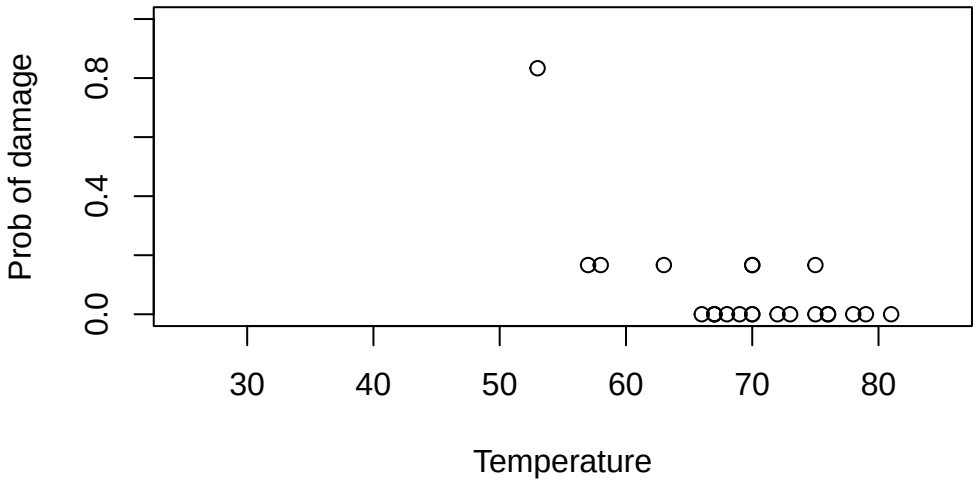
```
#data(orings)
orings %>% kbl() %>%
  kable_styling( full_width = F)
```

위의 자료에서 `temp`는 실험에서 적용된 온도(화씨 F)이고 `damage`는 각 실험마다 6개의 링중에서 손상된 개수를 나타낸다. 참고로 1986년 챌린져호가 발사될 때의 온도는 31F 였다.

먼저 그림을 통하여 온도의 변화에 따른 손상비율을 살펴보자.

```
plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim = c(0,1), xlab="Temperature", ylab="Prob of damage")
```

temp	damage
53	5
57	1
58	1
63	1
66	0
67	0
67	0
67	0
68	0
69	0
70	1
70	0
70	1
70	0
72	0
73	0
75	0
75	1
76	0
76	0
78	0
79	0
81	0



5.2.2. 강풍에 의한 나무 피해 자료

1999년 미국 Misnnesota주의 Boundary Waters Canoe Area Wilderness (BWCAW)에서 심한 폭풍으로 생긴 강한 바람에 의해 쓰러진 나무들에 대한 자료를 수집하였다. 이 자료는 Weisberg (2014) (R package `alr4`) 에 수록된 자료이다.

연구의 목적은 폭풍이 나무의 생존에 미치는 영향을 알아보는 것이다. 666 그루의 나무들에 대하여 나무가 바람에 의해 쓰러져 죽었는지 여부, 나무의 종, 나무의 지름, 폭풍의 국지적인 강도에 대한 자료를 수집하였다.

- `d`: Tree diameter, in cm
- `s` : Proportion of basal area killed for the four species balsam fir, cedar, paper birch and blue spruse, a measure of local severity of the storm.
- `spp` : Tree species, a factor with 9 levels
- `y` : 1 if the tree died, 0 if it survived

d	s	y	spp
9	0.0217509	0	balsam fir
14	0.0217509	0	balsam fir
18	0.0217509	0	balsam fir
23	0.0217509	0	balsam fir
9	0.0217509	0	balsam fir
16	0.0217509	0	balsam fir

d	s	y	spp
9	0.0242120	0	black spruce
11	0.0305947	0	black spruce
9	0.0305947	0	black spruce
9	0.0341815	0	black spruce
5	0.0341815	0	black spruce
8	0.0341815	0	black spruce

```
head(Blowdown) %>% kbl() %>%
  kable_styling( full_width = F)
```

반응변수 y 를 쓰러진 나무는 $y = 1$로하고 살아남은 나무를 $y = 0$으로 코딩하였다. 나무의 상태 y에 나무의 지름 d이 미치는 영향을 살펴보려고 한다.

수집된 자료중에서 나무의 종류가 black spruce인 자료만를 분석하기로 한다. 아래 코드는 black spruce인 자료만를 모아서 데이터프레임 BlowBS_raw 를 만드는 것이다.

```
BlowBS_raw <- Blowdown %>% dplyr::filter(spp=='black spruce')

dim(BlowBS_raw)
```

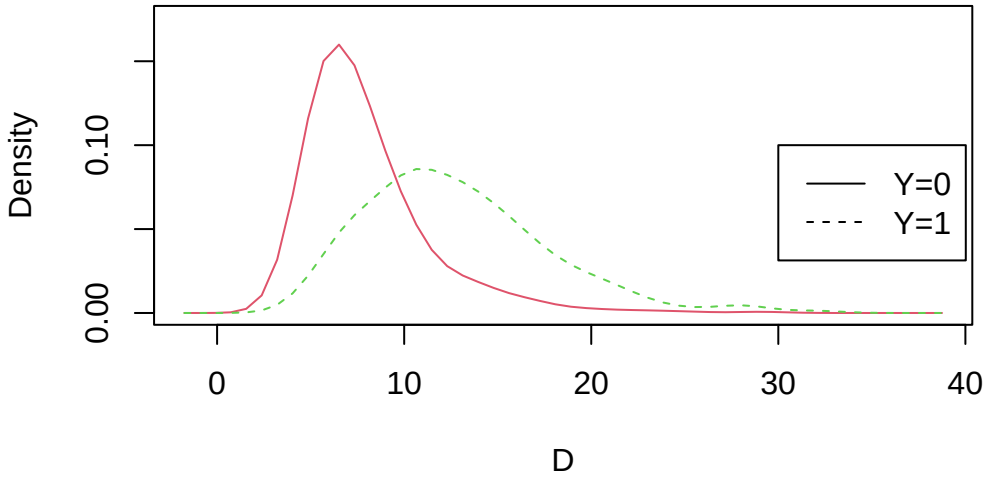
```
[1] 659   4
```

```
head(BlowBS_raw) %>% kbl() %>%
  kable_styling( full_width = F)
```

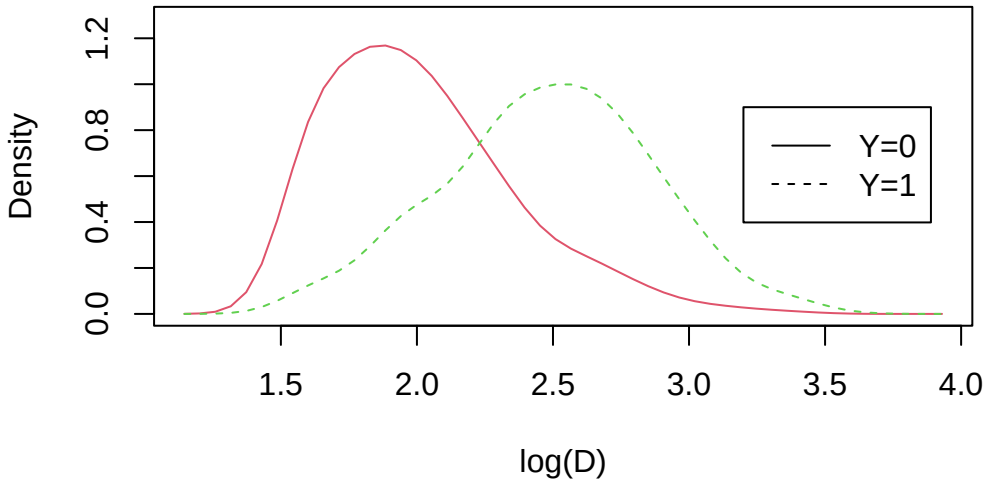
아래 그림은 쓰러진 나무와 살아남은 나무들의 지름의 분포를 비교한 것이다. 이를 통하여 지름이 큰 나무가 살아남을 확률이 더 커짐을 알수 있다.

```
sm.density.compare(BlowBS_raw$d, BlowBS_raw$y,lty=c(1,2), xlab="D")
legend(30,.1,legend=c("Y=0","Y=1"),lty=c(1,2))
```

d	died	m
5.0	6	88
5.5	1	2
6.0	6	91
6.5	1	1
7.0	17	90
7.5	1	1




```
sm.density.compare(log(BlowBS_raw$d), BlowBS_raw$y,lty=c(1,2), xlab="log(D)")
legend(3.2,.9,legend=c("Y=0","Y=1"),lty=c(1,2))
```



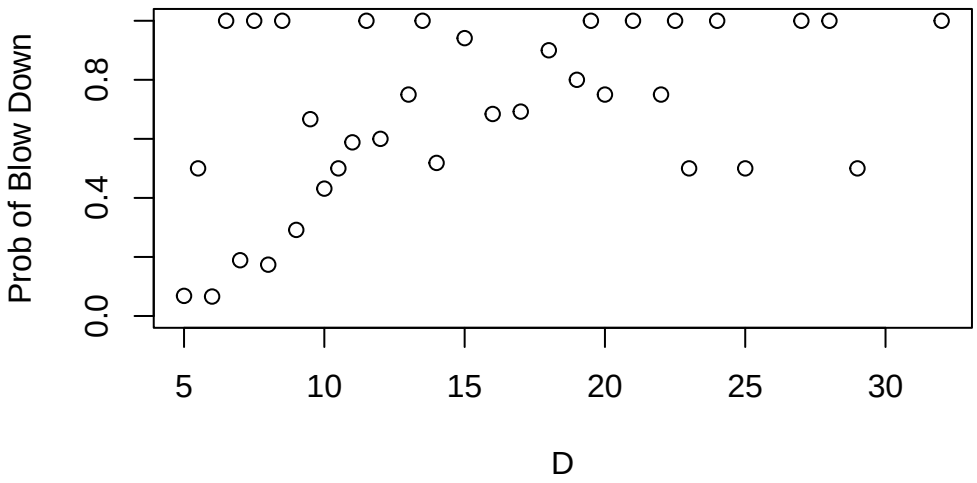
또한 R package `alr4` 에 수록된 데이터셋 `BlowBS` 는 위에서 본 O-rings 예제와 동일하게 `black spruce`인 자료만 모아서 전체 횟수와 성공의 횟수로 요약된 자료이다.

```
head(BlowBS) %>% kbl() %>%
  kable_styling( full_width = F)
```

- `d` :Tree diameter, in cm
- `died` : Number of trees of this value of d that died (blowdown)
- `m` : number of trees of this size class measured

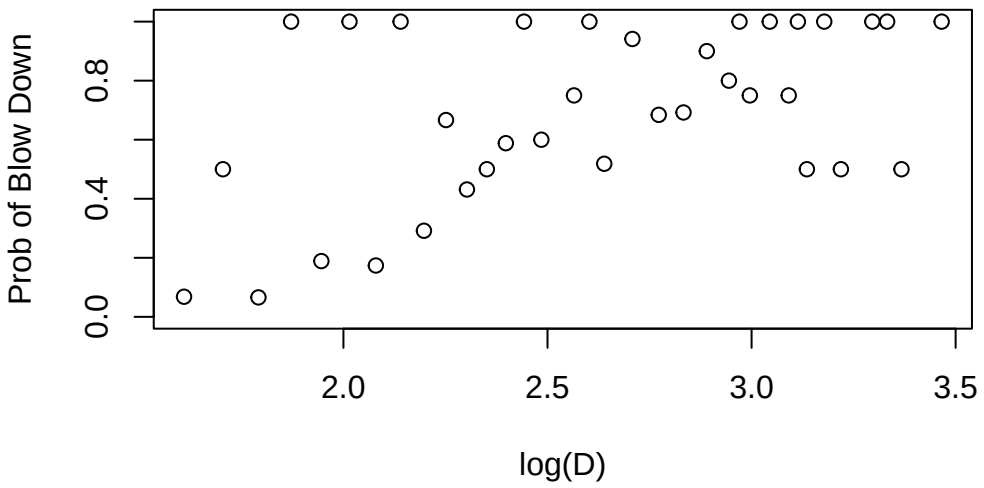
이제 데이터셋 `BlowBS` 를 이용하여 나무의 지름과 나무 피해의 관계에 대해서 살펴보자.

```
plot( died/m~ d, BlowBS, ylim = c(0,1), xlab="D", ylab="Prob of Blow Down")
```



나무의 지름 d을 $\log(d)$로 변환하여 예측변수(predictor) x로 하려고 한다.

```
plot( died/m~ I(log(d)), BlowBS, ylim = c(0,1), xlab="log(D)", ylab="Prob of Blow Down")
```



나무의 상태 y는 두 가지의 반응 결과를 가지는 이항변수이고 그 평균 $E(y)$는 0과 1사이의 값을 가지는 확률이지만 나무의 지름 $x = \log(D)$는 연속형 변수이다.

이러한 경우에 **회귀분석의 모형은 어떻게 세울까?**에 대하여 생각해보자.

$$0 \leq E(y|x) = p(x) \leq 1, \qquad -\infty < \beta_0 + \beta_1 x < \infty$$

5.3. 로지스틱 회귀모형

5.3.1. 이항변수와 연결함수

일반적으로 지금까지 배워온 회귀분석의 확률 모형에서는 반응변수 y는 연속형 확률변수이다. 따라서 예측변수 x의 값과 반응변수의 관계를 다음과 같은 회귀식으로 설명한다.

$$E(y|x) = \beta_0 + \beta_1 x \tag{5.1}$$

하지만 앞에서 살펴본 예제에서와 같이 반응변수의 값이 연속형 변수가 아니라 두 개의 가능한 결과만을 가지는 이항변수라면 위에서 주어진 회귀식은 적절하지 못하다.

$$y = \begin{cases} 1 & \text{with probabilty } p \\ 0 & \text{with probabilty } 1-p \end{cases}$$

왜냐하면 반응변수의 기대값이 0과 1사이의 확률로 나타나기 때문이다.

$$E(y|x) = 1 \cdot P(y=1|x) + 0 \cdot P(y=0|x) = P(y=1|x)$$

따라서 반응변수의 기대값의 범위와 예측변수가 있는 선형예측식(linear predictor) $\beta_0 + \beta_1 x$의 범위가 일치하지 않아서 선형회귀식 식 5.1 을 그대로 사용할 수 없다.

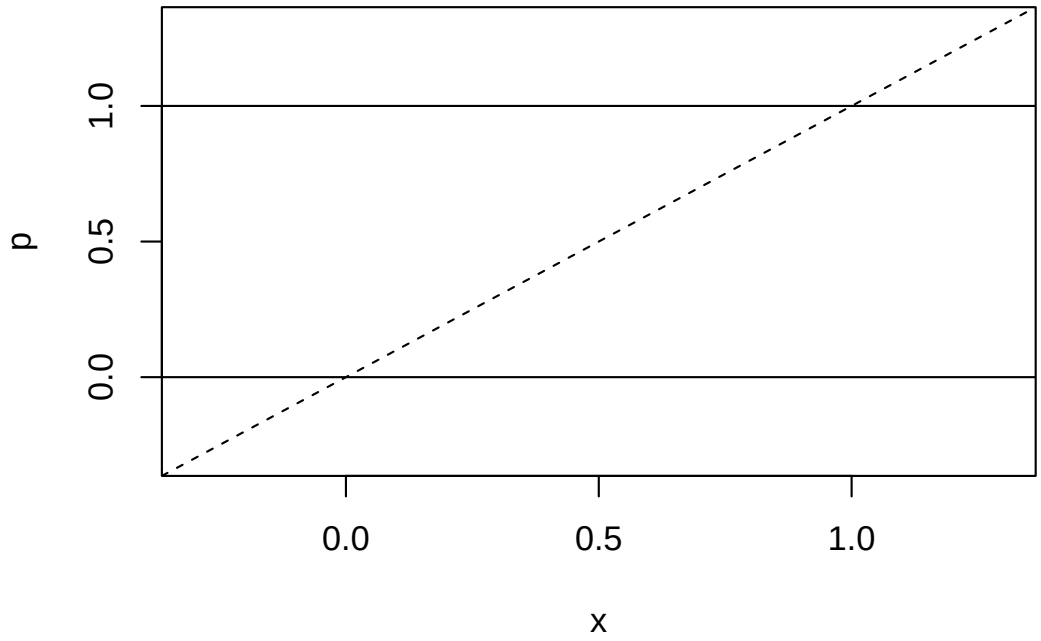


그림 5.1.: 범위의 불일치

위의 문제를 해결하기 위한 방법중의 하나는 다음과 같은 함수 m를 생각하여 변환된 선형예측식의 범위를 $[0,1]$로 만드는 것이다.

$$m : \Re \to [0,1] \quad \text{and} \quad m(x) \text{ is monotone function.}$$

따라서 다음과 같은 이항변수를 반응변수로 하는 새로운 회귀식을 만들 수있다.

$$E(y|x) = m(\beta_0 + \beta_1 x) \tag{5.2}$$

주로 쓰이는 변환함수로 다음과 같은 로지스틱 함수(logistic function)가 있다.

$$m(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{5.3}$$

반응변수가 베르누이 분포를 따를 때 위의 로지스틱함수를 사용하는 회귀식을 로지스틱 회귀식이라고 한다.

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1} \tag{5.4}$$

위의 로지스틱 회귀식을 다시 역으로 정리하면 다음과 같은 식을 얻을 수 있다.

$$\log\left[\frac{P(y = 1|x)}{1 - P(y = 1|x)}\right] = \log\frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x \tag{5.5}$$

식 식 5.5 에서 나타난 함수 g,

$$g(p) = \log\frac{p}{1 - p}$$

를 **로짓함수(logit function)** 라고 부르며 이는 로지스틱 함수의 역함수로서 0과 1 사이의 값을 가지는 확률을 실수 전체로 변환하는 함수로서 선형 예측식의 범위와 일치하게 한다.

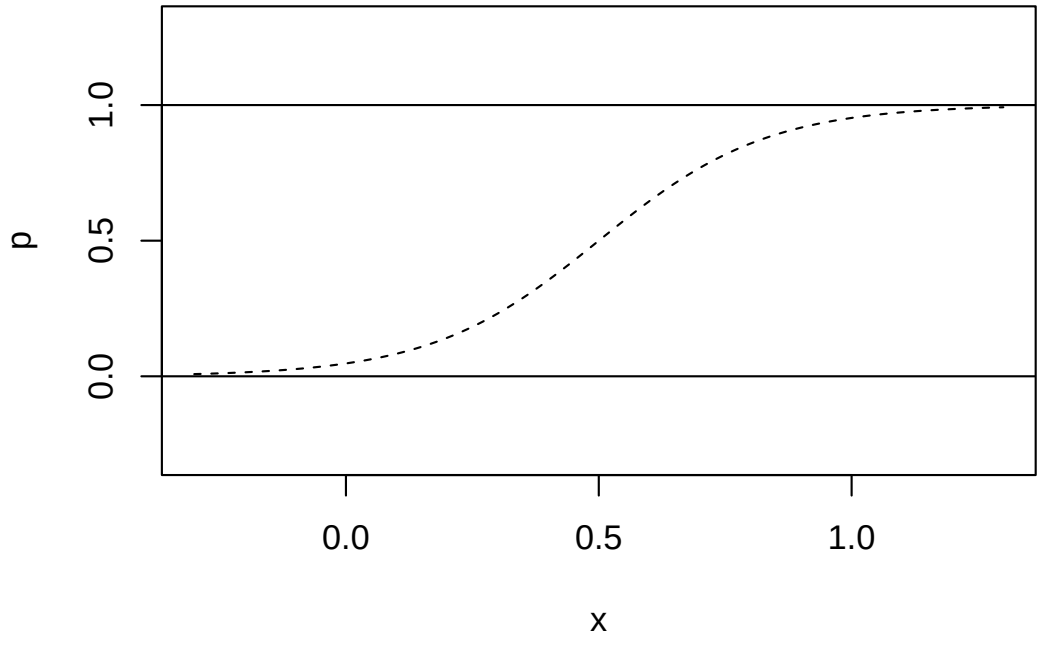


그림 5.2.: 로지스틱 연결함수

이렇게 관측값의 평균 (베르누이분포에서는 성공확률)과 선형예측식의 관계를 설정하는 함수를 **연결함수(link function)** 라고 하며 g 라고 표시한다.

$$g[E(y|x)] = g[p(x)] = \log\frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x \tag{5.6}$$

따라서 로짓함수는 연결함수의 하나이며 다른 종류의 연결함수도 생각할 수 있다. 예를 들어 $\Phi(x) = P(Z \leq x)$를 표준정규분포의 분포함수라 한다면 다음과 같은 연결함수를 생각할 수 있고 이를 probit 함수라고 부른다.

$$g[p(x)] = \Phi^{-1}(p(x)) = \beta_0 + \beta_1 x$$

만약 예측변수가 하나가 아닌 p개라면, 즉 예측변수 $\boldsymbol{x} = (x_1, x_2, \dots, x_p)^t$ 에 대한 로지스틱 회귀모형은 다음과 같이 확장할 수 있다.

$$\log\left[\frac{P(y=1|x)}{1-P(y=1|x)}\right] = \boldsymbol{x}^t\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p$$

5.3.2. 예제

위의 o-ring 예제(섹션 5.2.1)에서 성공($y=1$)은 O-ring이 손상된 경우이며 주어진 온도 x에서의 손상확률을 $p = (Y = 1|x) = p(x)$라고 하면 다음과 같은 로지스틱회귀식을 생각할 수 있다.

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

다음과 같은 함수 `glm`을 이용하여 위의 로지스틱 회귀식을 적합할 수 있다.

```
logit1 <- glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)
summary(logit1)
```

```
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
AIC: 33.675

Number of Fisher Scoring iterations: 6
```

```
coef(logit1)
```

```
(Intercept)        temp
 11.6629897  -0.2162337
```

위의 결과에서 회귀식 기울기의 추정치는 $\hat{\beta}$ = -0.2162337이다. 따라서 다음과 같은 회귀식을 얻을 수 있다.

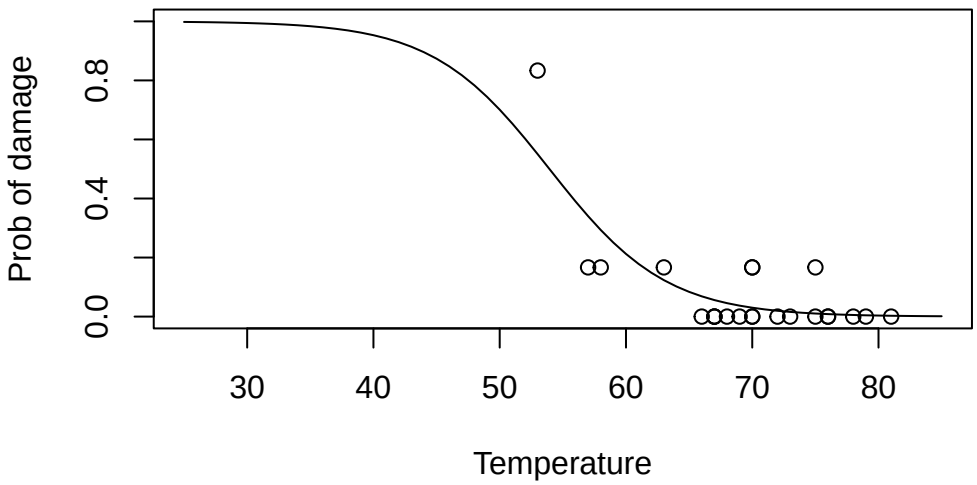
$$\log \frac{p(x)}{1-p(x)} = 11.6629897 + (-0.2162337)x$$

위의 회귀식을 이용하여 추정된 고장확률을 그림으로 그려보면 다음과 같다.

여기서 `ilogit`은 로짓함수의 역함수를 계산해주며 다음과 같이 주어진 회귀식을 이용하여 확률을 계산한다.

$$P(y=1|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} = \frac{1}{1+\exp[-(11.6629897+(-0.2162337)x)]}$$

```
plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim = c(0,1), xlab="Temperature", ylab="Prob of damage")
x <- seq(25,85,1)
lines(x,ilogit(coef(logit1)[1]+coef(logit1)[2]*x))
```



또한 온도가 31F인 경우 고장확률의 추정값은 $\hat{p} = 0.9930342$이며 R 에서 다음과 같이 계산한다.

```
x <- 31
ilogit(coef(logit1)[1]+coef(logit1)[2]*x)
```

```
(Intercept)
  0.9930342
```

이제 강풍에 의한 나무의 피해에 대한 예제(섹션 5.2.2)에 대하여 로지스틱 회귀식을 적합해보자.나무의 상태 y 가 이항 변수이고 나무의 지름 $x = \log(d)$을 예측변수로 하는 로지스틱회귀 모형을 고려하고 추정해보면 다음과 같은 회귀식을 얻는다

```
logit2 <- glm(y~ I(log(d)),family=binomial(),data=BlowBS_raw)
summary(logit2)
```

```
Call:
glm(formula = y ~ I(log(d)), family = binomial(), data = BlowBS_raw)

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.8925     0.6325  -12.48   <2e-16 ***
I(log(d))     3.2643     0.2761   11.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 856.21  on 658  degrees of freedom
Residual deviance: 655.24  on 657  degrees of freedom
AIC: 659.24

Number of Fisher Scoring iterations: 4
```

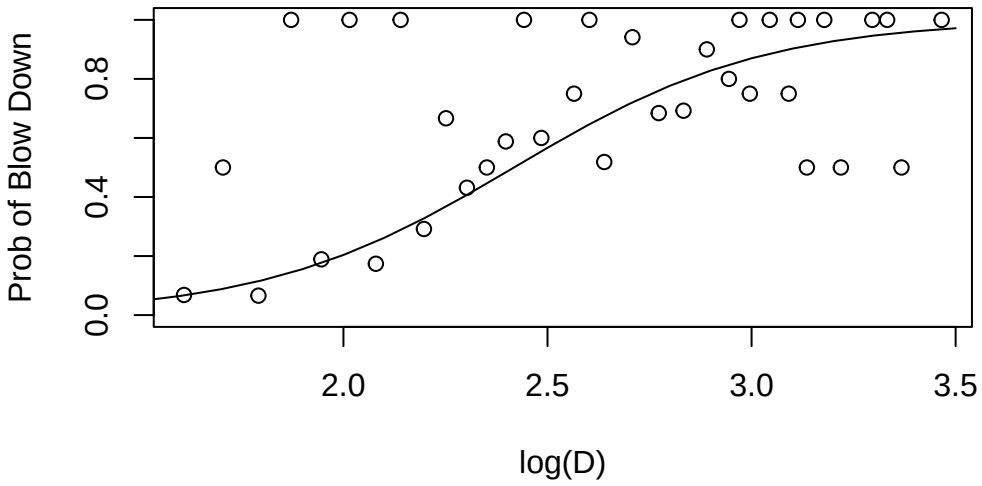
$$\log\left[\frac{P(y=1|x)}{1-P(y=1|x)}\right] = -7.892464 + (3.2642653)x$$

위의 회귀식을 이용하여 추정된 나무가 부러져서 피해를 입을 확률은 다음과 같다.

$$P(y=1|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} = \frac{1}{1+\exp[-(-7.892464+(3.2642653)x)]}$$

이제 나무의 지름에 변화에 대하여 나무가 부러져서 피해를 입을 확률의 추정값을 그림으로 그려보자.

```
plot( died/m~ I(log(d)), BlowBS, ylim = c(0,1), xlab="log(D)", ylab="Prob of Blow Down")
x <- seq(0,3.5,0.1)
lines(x,ilogit(coef(logit2)[1]+coef(logit2)[2]*x))
```



5.3.3. 회귀계수의 해석

일반적인 회귀분석의 모형 식 5.1 에서 계수 β_1은 기울기로서 예측변수 x의 단위가 1 증가할 때 반응변수의 평균이 β_1만큼 증가하는 것으로 해석할 수 있다. 하지만 로지스틱 회귀모형 식 5.4 에서는 이러한 해석을 할 수 없다.

로지스틱 회귀모형에서 기울기 β_1의 의미를 알아보기 위하여 예측변수 x 에 대한 베르누이 변수 y의 성공 확률 $P(y=1|x)$에 대한 오드(odd)를 정의하자.

$$odd(x) = \frac{P(y=1|x)}{1-P(y=1|x)}$$

이제 단순 로지스틱 회귀식 식 5.5 을 생각하고 예측변수 x를 0과 1의 값을 가지는 이항변수로 가정한다.

$x = 1$ 인 경우는

$$\frac{P(y=1|x=1)}{1-P(y=1|x=1)} = \exp(\beta_0 + \beta_1)$$

이며 $x = 0$ 인 경우는

$$\frac{P(y=1|x=0)}{1-P(y=1|x=0)} = \exp(\beta_0)$$

위에서 주어진 두 개의 오드, 즉 $x = 1$인 경우와 $x = 0$인 경우의 두 오드의 비(odd ratio)를 구하면 다음과 같다.

$$\frac{\frac{P(y=1|x=1)}{1-P(y=1|x=1)}}{\frac{P(y=1|x=0)}{1-P(y=1|x=0)}} = \exp(\beta_1)$$

이는 다시 쓰면

$$\frac{P(y=1|x=1)}{1-P(y=1|x=1)} = \exp(\beta_1)\frac{P(y=1|x=0)}{1-P(y=1|x=0)}$$

위의 식에서 볼 때 예측변수 x가 1 의 값을 가질 때 반응 변수의 오드가 예측변수 x가 0일 경우의 오드의 $\exp(\beta_1)$배로 변하는 것을 알 수 있다.

따라서 $\exp(\beta_1)$는 반응변수의 오드의 증가량으로 볼 수 있다. 이는 두 성공확률의 오즈비가 $\exp(\beta_1)$을 말한다. 위의 식에 로그를 취하면 다음과 같은 관계를 얻는다.

$$\log\left[\frac{P(y=1|x=1)}{1-P(y=1|x=1)} \Big/ \frac{P(y=1|x=0)}{1-P(y=1|x=0)}\right] = \beta_1$$

즉 오즈비의 로그값이 단순 로지스틱 회귀식에서 기울기 β_1으로 나타난다.

간단한 예제를 통하여 오즈비와 로지스틱 회귀의 기울기의 관계를 명확히 해보자. 100명의 사람들을 55세 이상의 사람($x = 1$)과 55세 미만의 사람($x = 0$)의 그룹으로 나누었을 때 각 그룹에서 만성심장질환(CHD)가 있는 사람($y = 1$)과 없는 사람($y = 0$)의 수가 표 표 5.1 에 주어져있다.

표 5.1.: 나이와 만성심장질환의 관계

CHD/나이	나이 ≥ 55 $(x = 1)$	나이 < 55 $(x = 0)$	합계
CHD 있음 $y = 1$	21	22	43
CHD 없음 $y = 0$	6	51	57
합계	27	73	100

여기서 나이에 대한 CHD 유무의 오즈비는 다음과 같이 계산된다.

$$\text{Odds Ratio} = \frac{\frac{21/27}{6/27}}{\frac{22/73}{51/73}} = \frac{(21)(51)}{(6)(22)} = 8.11$$

위의 표 표 5.1 의 자료를 이용하여 로지스틱회귀를 적합시키면 결과가 아래와 같고 회귀계수 β_1 의 추정값은 오즈비의 로그 값임을 알 수 있다.

$$\hat{\beta}_1 = \log(8.11) = 2.094$$

표 표 5.1 의 자료에 대하여 로지스틱 회귀모형은 다음과 같이 적합할 수 있다.

```r
yes <- c(21,22)
no  <- c(6,51)
x <- c (1,0)
m1 <- glm( cbind(yes,no) ~ x, family=binomial() )
summary(m1)
```

```
Call:
glm(formula = cbind(yes, no) ~ x, family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8408     0.2551  -3.296  0.00098 ***
x             2.0935     0.5285   3.961 7.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.8704e+01  on 1  degrees of freedom
Residual deviance: 1.4211e-14  on 0  degrees of freedom
AIC: 11.987

Number of Fisher Scoring iterations: 3
```

5.4. 추정과 검정

5.4.1. 이항분포와 가능도 함수

주어진 예측변수 x_i 에서 실행횟수가 m_i 인 이항분포$B(m_i, p(x_i))$를 생각하자. m_i 의 시행 중에 성공의 횟수가 y_i 라고 하면 y_i 의 확률밀도함수는 다음과 같다.

$$\binom{m_i}{y_i} p(x_i)^{y_i} [(1 - p(x_i)]^{m_i - y_i}$$

그리고 y_i의 평균과 분산은 다음과 같다.

$$E(y_i|x_i) = m_i p(x_i), \qquad Var(y_i|x_i) = m_i p(x_i)[1-p(x_i)], \quad i = 1, 2, \dots, n$$

이항분포를 위한 로지스틱 회귀방정식은 선형예측식과 성공의 확률의 관계를 다음과 같이 정한다.

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right] = \beta_0 + \beta_1 x_i$$

서로 독립인 관측값 $(y_1, y_2, \dots, y_n)$의 가능도함수(likelihood function) L은 이항분포들의 결합확률밀도함수와 같고 아래와 같이 주어지며

$$L = \prod_{i=1}^{n} f(y_i|p(x_i)) = \prod_{i=1}^{n} \left[\binom{m_i}{y_i}\right] p(x_i)_i^y (1-p(x_i))^{m-y_i}$$

로그가능도함수(log likelihood function) l 은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} l &= \log L \qquad (5.7)\\ &= \sum_i \log\binom{m_i}{y_i} + \sum_i y_i \log p(x_i) + \sum_i (m_i - y_i)\log[1-p(x_i)] \\ &= c(\boldsymbol{y}, \boldsymbol{m}) + \sum_i y_i \log\left[\frac{p(x_i)}{1-p(x_i)}\right] + \sum_i m_i \log[1-p(x_i)] \qquad (5.8) \end{aligned}$$

결론적으로

$$l(\boldsymbol{\mu}|\boldsymbol{y}) = \log L(\boldsymbol{\mu}|\boldsymbol{y}) = c(\boldsymbol{y}, \boldsymbol{m}) + \sum_i y_i \log\left[\frac{p(x_i)}{1-p(x_i)}\right] + \sum_i m_i \log[1-p(x_i)] \qquad (5.9)$$

위의 로그가능도함수에서 볼 수 있듯이 충분통계량인 성공의 횟수 y_i와 곱으로 나타내어진 함수가 로짓함수이며 이렇게 가능도함수에서 얻어진 결합함수를 자연 연결함수(natural link function)이라고 한다.

회귀계수의 추정량은 최대가능도 추정법을 이용하여 구할 수 있으며 로그 가능도 함수가 회귀계수에 대하여 비선형이므로 반복을 이용한 계산법에 의하여 추정량을 얻을 수 있다.

$$\max_{\boldsymbol{\beta}} \log L = \max_{\boldsymbol{\beta}} \sum_i y_i \log\left[\frac{p(x_i)}{1-p(x_i)}\right] + \sum_i m_i \log[1-p(x_i)]$$

5.4.2. 편차

선형모형에서 잔차제곱합(residual sum of square; SSE)에 대한 의미를 살펴보고 이를 일반화 선형모형에 확장하는 개념인 편차(deviance)의 정의를 알아보자.

먼저 다음과 같은 선형회귀식을 고려한다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots \beta_p x_{pi} + e_i, \quad i = 1, 2, \dots, n$$

여기서 오차항 e_i를 서로 독립이며 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정하고 (σ^2는 알고있다고 가정하자) 각 관측변수의 평균을 다음과 같이 μ_i로 하자.

$$\mu_i = E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

서로 독립인 관측변수 y_i의 분포는 정규분포를 따르므로

$$y_i \sim N(\mu_i, \sigma^2)$$

관측치 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$의 로그가능도함수는 다음과 같이 나타낼 수 있다.

$$l(\boldsymbol{\mu}|\boldsymbol{y}) = C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$$

예측변수 $x_1, x_2, \ldots, x_p$를 고려한 선형회귀모형에서 각 반응변수 평균의 예측식은 다음과 같다.

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi} \equiv \hat{y}_i$$

이 때 선형회귀모형의 로그가능도함수의 최대값는 다음과 같다.

$$\begin{aligned} l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) &= l_{regession}(\hat{\boldsymbol{\beta}}|y) \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}SSE \end{aligned}$$

이제 위의 선형회귀모형에서 예측변수 $x_1, x_2, \ldots, x_p$를 고려하지 않는 포화 모형을 생각해보자.

$$y_i = \beta_{0i} + e_i \quad \text{or} \quad E(y_i) = \beta_{0i}$$

이러한 포화 모형은 n개의 반응변수의 평균을 n개의 모수를 가진 모형으로 추정하는 것으로 위와 같은 모형을 포화 모형 (saturated model)이라고 한다. 포화 모형에서 모수 β_{0i}의 최소제곱 추정량(또는 최대가능도 추정량)는 관측값 y_i임을 쉽게 알수 있다.

$$\min_{\beta_{0i}} \sum_{i=1}^{n}(y_i - \beta_{0i})^2 \quad \Rightarrow \hat{\beta}_{0i} = y_i, \quad i = 1, 2, \ldots, n$$

포화 모형의 의미는 우리가 생각할 수 있는 모형 중에 가장 큰 모형으로 포화모형보다 큰 모형을 생각할 수 없다. 위에서 언급한 바와 같이 n개의 관측값에 대하여 모수의 수가 n개보다 큰 모형을 생각하면 유일한 모수의 추정이 불가능하다.

선형회귀모형에서 포화모형은 $\hat{\beta}_{0i} = y_i$이며 로그가능도함수의 최대값은 $l(\boldsymbol{y}|\boldsymbol{y})$로 표시하며 다음과 같다.

$$
\begin{aligned}
l(\boldsymbol{y}|\boldsymbol{y}) &= l_{saturated}(\hat{\boldsymbol{\beta}}_0|y) \\
&= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_{0i})^2 \\
&= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - y_i)^2 \\
&= C - \frac{n}{2}\log\sigma^2 + 0
\end{aligned}
$$

포화모형은 설정할 수 있는 최대의 모수를 가진 가장 큰 모형이므로 우리가 생각할 수 있는 모형 중에서 관측값을 예측하는 예측력은 가장 좋다는 것을 알 수 있다(하지만 과적합모형이다).

따라서 예측변수 $\boldsymbol{x}$들을 사용하는 선형회귀모형의 예측력이 포화모형이 가지는 예측력에 가까우면 좋은 모형이라고 생각할 수 있다. 반응변수의 평균을 예측하는 예측력은 로그가능도함수의 크기로서 나타낼 수 있다. 포화모형과 선형회귀모형의 로그가능도함수를 비교하면 포화모형의 로그가능도함수가 크다는 것을 알수 있고 (why?) 두 로그가능도함수의 의 차이를 비교하면 다음과 같다.

$$
l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) = l_{saturated}(\hat{\boldsymbol{\beta}}_0|y) - l_{regession}(\hat{\boldsymbol{\beta}}|y) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2\sigma^2}SSE
$$

위의 식을 가능도 함수의 비율로 다음과 같이 나타낼 수 있다.

$$
2\log\frac{L_{saturated}}{L_{regession}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{\sigma^2}SSE
$$

따라서 포화모형과 로그가능도함수의의 차이가 작다는 것은 선형회귀모형의 잔차제곱합(SSE)이 작다는 것을 의미한다. 보통 잔차제곱합이 작으면 선형회귀모형의 예측력이 좋은 모형이며 이는 선형회귀모형의 가능도함수가 포화모형의 가능도함수에 가깝다는 의미이다.

이렇게 모형의 예측능력을 평가하는 측도로서 편차(deviance)를 포화모형과 고려한 회귀모형의 로그가능도함수의 차이에 2를 곱한 양으로 정의한다. 따라서 편차는 작을 수록 좋다.

$$
deviance \equiv D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] = 2\log\frac{L_{saturated}}{L_{regession}} \tag{5.10}
$$

정규분포인 경우 편차는 다음과 같이 주어진다.

$$
D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] = \frac{1}{\sigma^2}SSE
$$

이제 이항분포들에서 나온 관측값에 대한 포화모형을 생각해 보자.

$$
y_i \sim B(m_i, p_i(x_i)), \quad i = 1, 2, \ldots, n
$$

위의 모형에서 포화모형은 어떤 모형일까? 포화모형은 n개의 관측변수의 평균, 여기서 $E(y_i/m_i) = p(\boldsymbol{x}_i)$를 n개의 관측치 y_i를 이용하여 추정한 모형으로서 각 성공확률은 해당하는 관측된 성공의 비율에 의해 추정된다. 즉,

$$\hat{p}(x_i) = \frac{y_i}{m_i}$$

이러한 경우의 로그가능도함수의 값은 다음과 같이 주어진다.

$$\begin{aligned} l(\boldsymbol{y}|\boldsymbol{y}) &= l_{saturated} \\ &= \sum_i \log \binom{m_i}{y_i} + \sum_i y_i \log \hat{p}(x_i) + \sum_i (m_i - y_i) \log(1 - \hat{p}(x_i)) \\ &= \sum_i \log \binom{m_i}{y_i} + \sum_i y_i \log \frac{y_i}{m_i} + \sum_i (m_i - y_i) \log(1 - \frac{y_i}{m_i}) \end{aligned}$$

따라서 위에서 주어진 포화함수의 로그가능도함수에서 로지스틱회귀식의 로그가능도함수 식 5.9 를 빼고 2를 곱해서 편차를 정의할 수 있다.

$$\begin{aligned} D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) &= 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] \\ &= 2(l_{saturated} - l_{regession}) \\ &= 2\left[\sum_i y_i \log \frac{y_i}{m_i} + \sum_i (m_i - y_i) \log(1 - \frac{y_i}{m_i}) - \sum_i y_i \log \hat{p}(x_i) - \sum_i (m_i - y_i) \log(1 - \hat{p}(x_i))\right] \\ &= 2\left[\sum_i y_i \log \frac{y_i}{m_i \hat{p}(x_i)} + \sum_i (m_i - y_i) \log \frac{1 - y_i/m_i}{1 - \hat{p}(x_i)}\right] \\ &= 2\left[\sum_i y_i \log \frac{y_i}{m_i \hat{p}(x_i)} + \sum_i (m_i - y_i) \log \frac{m_i - y_i}{m_i - m_i \hat{p}(x_i)}\right] \\ &= 2\left[\sum_i y_i \log \frac{y_i}{\hat{y}_i} + \sum_i (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{y}_i}\right] \end{aligned}$$

위에서 $\hat{y}_i = m_i \hat{p}(x_i)$으로 로지스틱 회귀에서 성공의 횟수의 평균에 대한 예측값이다.

위의 논의에서 알 수 있듯이 로지스틱 회귀에서의 편차는 선형회귀 분석에서 잔차 제곱합 SSE의 의미로 해석할 수 있으며 작을 수록 모형의 예측력이 좋다는 것을 알 수 있다.

! 편차의 점근적 분포

편차는 표본의 개수 m_i가 충분히 크고 회귀식이 옳다는 가정 하에서 자유도가 $n - p$ 인 카이제곱분포를 따른다. 여기서 p는 회귀계수 벡터 $\boldsymbol{\beta}$의 크기이다.

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}$$

여기서 $n = \sum_i m_i$ 이고 p 는 회귀계수의 갯수이다.

정규분포와 이항분포의 편차를 비교하면 정규분포의 편차에는 산포를 나타내는 모수 σ^2 이 포함되어 있지만 이항분포의 편차에는 다른 모수가 나타나지 않는다. 식 식 5.10 에서 주어진 편차를 척도 모수(scaled parameter) 또는 산포 모수(dispersion parameter) ϕ 를 곱해준 값을 척도화 편차(scaled deviance) $D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ 라고 부른다.

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \phi D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) \tag{5.11}$$

정규분포에서 산포 모수가 분산 $\phi = \sigma^2$ 이므로 척도화 편차는 잔차제곱합 $D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = SSE$ 가 되며 이항분포에서는 편차와 척도화 편차가 같다.

이제 다시 o-ring 의 예제(섹션 5.2.1)에 대한 로지스틱 회귀식의 | 결과를 보자.

```
summary(logit1)
```

```
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
AIC: 33.675

Number of Fisher Scoring iterations: 6
```

마지막에 `Residual deviance` 는 다음과 같이 절편과 하나의 예측 변수(온도)를 포함한 회귀식에 대한 편차의 값과 자유도를 나타내는 것이다.

$$\log\left[\frac{P(y=1|x)}{1-P(y=1|x)}\right] = \beta_0 + \beta_1 x \quad \rightarrow D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 16.9122785$$

```
deviance(logit1)
```

```
[1] 16.91228
```

```
df.residual(logit1)
```

```
[1] 21
```

위에서 언급하였듯이 모형이 옳다는 가정하에서 편차 $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ 는 자유도가 21 인 χ^2-분포를 따르므로 아래에서 구한 $P(\chi^2 \geq 16.9122785)$ 가 크게 나오는 것은 모형이 적절하다는 의미이다.

```
pchisq(deviance(logit1), df.residual(logit1),lower=FALSE)
```

```
[1] 0.7164099
```

또한 `Null deviance` 는 다음과 같이 절편만 포함한 회귀식에 대한 편차의 값과 자유도를 나타내는 것이다.

$$\log\left[\frac{P(y=1|x)}{1-P(y=1|x)}\right] = \beta_0$$

5.4.3. 검정과 모형의 선택

로지스틱 회귀모형에서 하나의 회귀 계수가 유의한 지에 대한 다음 검정은

$$H_0 : \beta_i = 0 \quad vs. \quad H_1 : \beta_i \neq 0$$

다음과 같은 표준화된 통계량을 이용하여 정규분포 검정을적용할 수 있다.

$$z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

또한 회귀 계수 β_i 에 대한 $(1-\alpha)100\%$ 신뢰구간은 다음과 같이 구할 수 있다.

$$[\hat{\beta}_i - z_{\alpha/2}SE(\hat{\beta}_i), \hat{\beta}_i + z_{\alpha/2}SE(\hat{\beta}_i)]$$

위의 o-ring 예제(섹션 5.2.1)에서 온도에 대한 회귀 계수의 95% 신뢰구간은 다음과 같이 `MASS` 패키지의 함수 `confint`를 이용해서 구한다.

```
confint(logit1)
```

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

```
                2.5 %    97.5 %
(Intercept)  5.575195 18.737598
temp        -0.332657 -0.120179
```

이제 회귀모형에서 고려하는 다음과 같은 가설 검정을 고려해 보자.

$$H_0 : \text{ reduced model} \quad vs. \quad H_1 : \text{full model} \tag{5.12}$$

로지스틱 회귀에서 편차(deviance)의 차이를 이용하여 두 개의 모형 중 하나를 선택하는 방법을 살펴보자.

다음과 같이 서로 다른 모형과 그에 대한 가설을 생각해 보자.

$$H_0 : g[p(x)] = \boldsymbol{x}_1^t\boldsymbol{\beta}_1 \quad \text{vs.} \quad H_1 : g[p(x)] = \boldsymbol{x}_1{}^t\boldsymbol{\beta}_1 + \boldsymbol{x}_2^t\boldsymbol{\beta}_2$$

가설을 달리 표현하면 다음과 같이 쓸 수 있다.

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 \neq 0$$

위의 가설에서 $dim(\beta_1) = p$, $dim(\beta_2) = q$ 라고 하면 대립가설 H_1의 모형이 귀무가설 H_0의 모형보다 큰 모형이다.

만약 축소모형(reduced moldel, H_0)에 대한 잔차 편차(Residual Deviance)를 D_0 라고 하고 큰 모형(full model, H_1)에 대한 잔차 편차를 D_1 라고 하면 귀무가설이 참인 경우 두 편차의 차이 $D_0 - D_1$ 은 근사적으로 자유도가 q 인 카이제곱분포를 따른다. 여기서 자유도 q 는 두 모형의 회귀계수의 갯수 차이 $(p+q)-p=q$이다. 이러한 두 편차의 차이의 점근적 분포 가능도비검정 이론에 의하여 유도할 수 있다.

$$D = D_0 - D_1$$

두 모형에 대한 모수의 개수의 차이 q인 경우 귀무가설이 참일 때 deviance의 차인 $D = D_0 - D_1$ 통계량은 자유도가 q인 χ^2-분포를 따른다. 따라서 유의수준 α에서 D 통계량이 $\chi^2_\alpha(q)$보다 크면 귀무가설을 기각한다.

편차 차이를 이용하여 검정하는방법은 가능도비 검정(likelihood ratio test)과 동일한 검정이다.

다음 자료는27명의 암환자들에 대한 암의 호전(remission of cancer)이며 종속변수 y는 **remiss**로 1이면 암이 호전되었다는 표시이다. 나머지 6개의 변수는 환자의 특성을 나타내는 독립변수이다.

```
cancer <- read.table("remission.txt",header=T, sep="")
cancer %>% kbl() %>%
  kable_styling( full_width = F)
```

다음은 6개의 독립변수를 모두 적합한 완전모형(full model)의 추정 결과이다.

```
logit3_L <- glm(remiss~cell + smear + infil + li + blast + temp,family=binomial,data=cancer)
summary(logit3_L)
```

```
Call:
glm(formula = remiss ~ cell + smear + infil + li + blast + temp,
    family = binomial, data = cancer)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  58.0385    71.2364   0.815   0.4152
cell         24.6615    47.8377   0.516   0.6062
smear        19.2936    57.9500   0.333   0.7392
infil       -19.6013    61.6815  -0.318   0.7507
li            3.8960     2.3371   1.667   0.0955 .
blast         0.1511     2.2786   0.066   0.9471
temp        -87.4339    67.5735  -1.294   0.1957
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

id	remiss	cell	smear	infil	li	blast	temp
1	1	0.80	0.83	0.66	1.9	1.100	0.996
2	1	0.90	0.36	0.32	1.4	0.740	0.992
3	0	0.80	0.88	0.70	0.8	0.176	0.982
4	0	1.00	0.87	0.87	0.7	1.053	0.986
5	1	0.90	0.75	0.68	1.3	0.519	0.980
6	0	1.00	0.65	0.65	0.6	0.519	0.982
7	1	0.95	0.97	0.92	1.0	1.230	0.992
8	0	0.95	0.87	0.83	1.9	1.354	1.020
9	0	1.00	0.45	0.45	0.8	0.322	0.999
10	0	0.95	0.36	0.34	0.5	0.000	1.038
11	0	0.85	0.39	0.33	0.7	0.279	0.988
12	0	0.70	0.76	0.53	1.2	0.146	0.982
13	0	0.80	0.46	0.37	0.4	0.380	1.006
14	0	0.20	0.39	0.08	0.8	0.114	0.990
15	0	1.00	0.90	0.90	1.1	1.037	0.990
16	1	1.00	0.84	0.84	1.9	2.064	1.020
17	0	0.65	0.42	0.27	0.5	0.114	1.014
18	0	1.00	0.75	0.75	1.0	1.322	1.004
19	0	0.50	0.44	0.22	0.6	0.114	0.990
20	1	1.00	0.63	0.63	1.1	1.072	0.986
21	0	1.00	0.33	0.33	0.4	0.176	1.010
22	0	0.90	0.93	0.84	0.6	1.591	1.020
23	1	1.00	0.58	0.58	1.0	0.531	1.002
24	0	0.95	0.32	0.30	1.6	0.886	0.988
25	1	1.00	0.60	0.60	1.7	0.964	0.990
26	1	1.00	0.69	0.69	0.9	0.398	0.986
27	0	1.00	0.73	0.73	0.7	0.398	0.986

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 21.751  on 20  degrees of freedom
AIC: 35.751

Number of Fisher Scoring iterations: 8
```

위의 회귀분석 결과를 보면 변수들 중 세 개의 변수 `li`, `temp` , `cell` 만 포함한 축소된 모형(reduced model)을 생각해보자.

```
logit3_S <- glm(remiss~cell + li + temp,family=binomial,data=cancer)
summary(logit3_S)
```

```
Call:
glm(formula = remiss ~ cell + li + temp, family = binomial, data = cancer)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   67.634     56.888   1.189   0.2345
cell           9.652      7.751   1.245   0.2130
li             3.867      1.778   2.175   0.0297 *
```

```
temp         -82.074     61.712  -1.330   0.1835
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 21.953  on 23  degrees of freedom
AIC: 29.953

Number of Fisher Scoring iterations: 7
```

위에서 적합한 두개의 모형이 유의한 차이가 있는지 위에서 논의한 편차의 차이를 이용하여 검정해 보자. 다음과 같은 가설을 검정하는 검정은 다음과 같이 anova함수로 수행할 수 있다.

앞에서 두 모형을 적합한 결과에서 잔차 편차(Residual deviance)의 차이를 이용하여 검정을 실시하였다.

```
anova(logit3_S, logit3_L,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: remiss ~ cell + li + temp
Model 2: remiss ~ cell + smear + infil + li + blast + temp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        23     21.953
2        20     21.751  3  0.20272   0.9771
```

위의 anova 함수의 결과를 보면 p-vlaue가 0.9771497로 매우 크며 귀무가설 H_0를 기각하지 못한다. 따라서 3개의 독립변수만 가진 축소된 모형을 선택할 수 있다.

다시 하나의 독립변수 li 만 가지는 축소모형을 고려해 보자.

```
logit3_S2 <- glm(remiss~li,family=binomial,data=cancer)
summary(logit3_S2)
```

```
Call:
glm(formula = remiss ~ li, family = binomial, data = cancer)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.777      1.379  -2.740  0.00615 **
li             2.897      1.187   2.441  0.01464 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom
AIC: 30.073

Number of Fisher Scoring iterations: 4
```

완전모형과의 차이에 대한검정을 실시하며 다음과 같은 결과가 나오고 cencer remission 자료는 사실상 변수 `li`만으로도 충분히 설명할 수 있다는 결론이다.

```
anova(logit3_S2, logit3_L,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: remiss ~ li
Model 2: remiss ~ cell + smear + infil + li + blast + temp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        25     26.073
2        20     21.751  5   4.3223    0.504
```

모형을 선택하는 경우 AIC(Akaike Information Criteria)를 자주 이용한다. AIC는 다음과 같이 정의되며

$$AIC = -2 * log(\text{Likelihood}) + 2 * (\text{number of parameter})$$

모형의 AIC값이 작을수록 좋은 모형이다.

```
AIC(logit3_L, logit3_S, logit3_S2)
```

```
          df      AIC
logit3_L   7 35.75065
logit3_S   4 29.95337
logit3_S2  2 30.07296
```

위에서 나타난 AIC로 모형을 선택한다면 독립변수가 3개(`li`, `temp` , `cell`)인 모형이 가장 좋은 모형이다. AIC는 모형에 위에서 주어진 두 번째 항에서 독립변수의 수(number of parameter)를 더해주므로 모형이 복잡할수록 그 값이 증가하여 모형의 선택에서 벌칙(penalty)를 주는 것으로 이해할 수 있다.

AIC는 로그우도함수와 모형의 자유도로 직접 계산할 수 있다.

```
AIC(logit3_S)
```

```
[1] 29.95337
```

```
# value of log likelihood
as.numeric(logLik(logit3_S))
```

```
[1] -10.97668
```

```
#number of parameter = 3+1
attr(logLik(logit3_S),"df")
```

```
[1] 4
```

```
#AIC
-2*as.numeric(logLik(logit3_S))+2*attr(logLik(logit3_S),"df")
```

```
[1] 29.95337
```

위에서 모형의 자유도는 독립변수의 개수 3개와 절편 1개를 더해서 4개가 된다.

5.4.4. 적합도 검정

모형의 적합도를 측정하는 통계량으로 편차(deviance)뿐만 아니라 χ^2-통계량도 매우 유용하다. χ^2-통계량은 다음과 같이 주어진다.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{p}(x_i))^2}{m_i\hat{p}(x_i)[1-\hat{p}(x_i)]} \tag{5.13}$$

위의 식은 다음과 같은 표준화된 피어슨 잔차 r_i 의 제곱합과 같다.

$$r_i = \frac{y_i - m_i\hat{p}(x_i)}{\sqrt{var(\hat{y}_i)}}$$

자료의 수가 많으면 χ^2-통계량은 χ^2 분포를 따른다.

χ^2-통계량은 다음과 같이 계산할 수 있다.

```
sum(residuals(logit1,type="pearson")^2)
```

```
[1] 28.06738
```

5.5. 과산포

5.5.1. 과산포의 개요

로지스틱 회귀식을 적합한 뒤에 주어진 모형이 적절한 모형이라고 판단되지만 잔차 편차(residual devince) D가 기대 이상으로 너무 커서 서로 상충되는 결론이 나오면 다음과 같은 원인을 의심할 수 있다.

- 중요한 예측변수 x가 모형에 포함되지 않았거나 예측함수 $\eta = \boldsymbol{x}^t\boldsymbol{\beta}$가 잘못 설정된 경우
- 이상점(outlier)이 있는 경우
- 반복수의 수가 매우 적은 경우

- 분포의 가정이 맞지 않는 경우 - 독립이 아닌 경우, 군집효과(cluster effect)

반응변수 y가 이항분포 $B(m,p)$ 를 따르면 그 평균과 분산은 다음과 같이 주어진다.

$$E(y) = mp \quad \text{and} \quad Var(y) = mp(1-p)$$

자료의 확률 구조가 이항분포의 평균과 분산에 대한 가정과 맞지 않는 경우를 분산의 크기에 따라 과산포(overdispersion) 또는 underdispersion이라고 한다.

- overdispersion if $Var(y) > mp(1-p)$
- underdispersion if $Var(y) < mp(1-p)$

예를 들어 모집단이 몇 개의 군집으로 이루져 있다면 군집효과로 발생한 성질때문에 관측치의 분산이 이항분포의 분산보다 큰 경우가 생기게 된다. 이러한 경우를 overdispersion이라고 한다.

군집 효과는 다음과 같은 이유로 발생할 수 있다.

- 군집간 의 분포가 서로 다른 경우
- 관측값이 서로 독립이 아닌 경우

표본의 개수를 n이라고 하고 군집의 크기를 k 라 하면 군집의 개수는 $l = n/k$ 이 된다.

- i번째 군집에서 성공의 회수 z_i는 이항분포 $B(k, p_i)$를 따른다고 하자.
- 이 때, 성공의 확률 p_i를 평균이 $E(p_i) = p$이고 분산이 $Var(p_i) = \tau^2 p(1-p)$ 인 확률변수라고 가정하자.

  이제 총 성공의 횟수 $y = z_1 + \cdots + z_k$의 분포를 살펴보면

$$E(y) = \sum_i E(Z_i) = \sum_i kp = mp$$

로서 평균은 보통의 경우와 같지만 분산은 다음과 같이 overidspersion이 나타난다.

$$\begin{aligned}
Var(y) &= \sum_i Var(z_i) \\
&= \sum_i [E(Var(z_i|p_i)) + Var(E(z_i|p_i)] \\
&= \sum_i [Ekp_i(1-p_i) + Var(kp_i)] \\
&= \sum_i [kE(p_i) - kE(p_i^2) + k^2\tau^2 p(1-p)] \\
&= \sum_i [kp - k\{\tau^2 p(1-p) + p^2\} + k^2\tau^2 p(1-p)] \\
&= \sum_i [kp - k\{\tau^2 p(1-p) + p^2\} + k^2\tau^2 p(1-p)] \\
&= [1 + \tau^2(k-1)]kp(1-p)
\end{aligned}$$

즉 $1 + \tau^2(k-1)$의 값이 1보다 크기 때문에 반응변수가 독립인 경우의 분산 $kp(1-p)$보다 커진다. $1 + \tau^2(k-1)$ 의 값을 산포모수(dispersion parameter) 라고 부른다.

! 산포모수

이항변수에 대한 로지스틱회위에서 다음과 같이 분산이 이항분포의 이론적 분산 $kp(1-p)$ 보다 크게 크게 나타나는 경우 그 계수를 산포모수(dispersion parameter, σ^2) 이라고 부른다.

$$Var(y) = \sigma^2 mp(1-p) > kp(1-p)$$

산포모수(σ^2)는 과산포가 없으면 1 이고 과산포가 존재하면 1 보다 큰 값을 가진다.

산포모수는 식 식 5.13 에 주어진 χ^2-통계량에 의해 추정될 수 있다.

$$\hat{\sigma}^2 = \chi^2/(n-p)$$

주의할 점은 overidspersion이 있다 하더라도 회귀계수의 추정치 β의 추정에는 영향을 미치지 않는다.

위와 같이 overidspersion이 있다고 판단되는 경우에는 추가로 산포모수(dispersion parameter) σ^2를 고려하고 이를 추정하여 회귀계수 β의 분산 계산 시 다음과 같이 추가해서 계산한다.

$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (X'\hat{W}X)^{-1}$$

5.5.2. 예제

부교재 Faraway (2016) 의 2.11 절에 소개된 자료를 이용하여 과산포에 대한 예제를 살펴보자.

데이터 `trouegg` 는 5개의 장소와 4개의 시번에서 관측한 송어(trout) 알의 생존에 대한 자료이다. 데이터를 구성하는 변수의 이름과 의미는 다음과 같다.

- `survive`: the number of surviving eggs
- `total`: the number of eggs in the box
- `location` : the location in the stream with levels 1 2 3 4 5
- `period` : the number of weeks after placement that the box was withdrawn levels 4 7 8 11

다음은 데이터 `trouegg` 이다.

```
troutegg %>% dplyr::arrange(location, period) %>% dplyr::relocate(`location`, `period`, `survive`, `tot
 kbl() %>%
  kable_styling( full_width = F)
```

```
summary(troutegg)
```

```
    survive           total        location period
 Min.   :  0.00   Min.   : 86.00   1:4      4 :5
 1st Qu.: 74.50   1st Qu.: 96.75   2:4      7 :5
 Median : 90.00   Median :107.00   3:4      8 :5
 Mean   : 81.35   Mean   :111.30   4:4      11:5
 3rd Qu.:104.00   3rd Qu.:123.50   5:4
 Max.   :141.00   Max.   :155.00
```

	location	period	survive	total
1	1	4	89	94
6	1	7	94	98
11	1	8	77	86
16	1	11	141	155
2	2	4	106	108
7	2	7	91	106
12	2	8	87	96
17	2	11	104	122
3	3	4	119	123
8	3	7	100	130
13	3	8	88	119
18	3	11	91	125
4	4	4	104	104
9	4	7	80	97
14	4	8	67	99
19	4	11	111	132
5	5	4	49	93
10	5	7	11	113
15	5	8	18	88
20	5	11	0	138

이제 자료 troutegg 에 로지스틱 회귀식을 다음과 같이 적합해보자.

```
bmod <- glm(cbind(survive,total-survive) ~      location+period,          family=binomial,troutegg)
summary(bmod)
```

```
Call:
glm(formula = cbind(survive, total - survive) ~ location + period,
    family = binomial, data = troutegg)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.6358     0.2813  16.479  < 2e-16 ***
location2    -0.4168     0.2461  -1.694   0.0903 .
location3    -1.2421     0.2194  -5.660 1.51e-08 ***
location4    -0.9509     0.2288  -4.157 3.23e-05 ***
location5    -4.6138     0.2502 -18.439  < 2e-16 ***
period7      -2.1702     0.2384  -9.103  < 2e-16 ***
period8      -2.3256     0.2429  -9.573  < 2e-16 ***
period11     -2.4500     0.2341 -10.466  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1021.469  on 19  degrees of freedom
Residual deviance:   64.495  on 12  degrees of freedom
AIC: 157.03
```

```
Number of Fisher Scoring iterations: 5
```

앞의 예제와 같이 결과의 잔차 편차가 χ^2-분포를 기준으로 얼마나 큰지 살펴보자.

```
pchisq(deviance(bmod), df.residual(bmod),lower=FALSE)
```

```
[1] 3.379416e-09
```

잔차 편차가 매우 큰 것을 알 수 있으며 산포모수 σ^2 는 식 식 5.13 에 주어진 것 처럼 다음과 같이 추정될 수 있다.

```
sigma2 <- sum(residuals(bmod,type="pearson")^2) /12
sigma2
```

```
[1] 5.330322
```

위에서 구한 산포 모수가 크다는 것은 다음과 같이 o-ring 자료에 대한 로지스틱 회귀의 산포 모수와 비교하면 알 수 있다.

```
sigma2 <- sum(residuals(logit1,type="pearson")^2) /21
sigma2
```

```
[1] 1.336542
```

6. 포아송 회귀모형

6.1. 필요한 패키지

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
library(alr4)
library(MASS)
library(knitr)
library(kableExtra)
library(pscl)
library(here)
```

6.2. 포아송 분포

반응변수 y가 어떤 사건이 일어난 횟수(count)라면 주로 포아송분포를 확률 모형으로 사용한다.

$$P(Y=y)=f(y|\mu)=\frac{e^{-\mu}\mu^{y}}{y!}, \quad y=0,1,2,... \tag{6.1}$$

포아송 분포는 다음과 같은 중요한 특성을 가지고 있다.

- 만약에 어떤 사건이 일어난 횟수가 몇 가지 가능한 수들 중에 하나라면 (예: $0 \leq y \leq M$) 포아송분포를 이항분포의 근사(approximation)로 생각할 수 있다. 만약 n이크고 성공확률 p가 작으면 이항분포는 평균이 $\mu=np$인 포아송 분포와 매우 가깝기 때문에 가능한 횟수가 제한되었다 하더라도 포아송 분포를 적용할 수 있다.
- 사건의 일어난 횟수가 주어진 시간의 길이에 비례하고 다른 사건과 독립이면 포아송 분포를 따른다. 또한 포아송 분포는 두 개의 사건이 일어날 때 시간 간격이 지수분포(exponential distribution)을 따른다면 주어진 시간 간격동안 일어난 사건의 횟수는 포아송 분포를 따른다.
- y_i가 서로 독립이고 평균이 μ_i인 포아송분포를 따른다면 합 $\sum_i y_i$는 평균이 $\sum_i \mu_i$인 포아송분포를 따른다

6.3. 포아송 회귀모형

이러한 포아송 분포에서 나온 반응변수(횟수 y)에 대하여 설명변수 x의 영향에 대한 회귀분석을 포아송 회귀모형이라고 한다.

포아송 분포의 평균 μ는 양의 실수이고 선형예측식 $\eta = \boldsymbol{x}^t\boldsymbol{\beta}$의 범위는 실수이기 때문에 로그함수를 연결함수(link function)으로 이용하여 회귀식을 세운다.

$$\log E(y|\boldsymbol{x}_i) = \log \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{6.2}$$

포아송 회귀모형에서 회귀 계수의 의미를 알아보기 위하여 다음과 같이 하나의 설명변수만 있으며 그 값이 $x = 0$ 과 $x = 1$인 경우에 회귀식을 보자.

$$\log E(y|x = 1) = \log \mu(x = 1) = \beta_0 + \beta_1, \quad \log E(y|x = 2) = \log \mu(x = 2) = \beta_0 + 2\beta_1$$

따라서 다음과 같은 식이 성립하므로 설명변수 x 가 1 단위 증가하면 반응변수의 평균은 $\exp(\beta_1)$ 배 증가한다는 것을 알 수 있다.

$$\frac{E(y|x = 2)}{E(y|x = 1)} = \frac{\mu(x = 2)}{\mu(x = 1)} = \frac{\exp(\beta_0 + 2\beta_1)}{\exp(\beta_0 + \beta_1)} = \exp(\beta_1)$$

사실 포아송 분포의 로그가능도함수에서 로그함수가 자연 연결함수임을 쉽게 알 수 있다. 즉, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$를 서로 독립이고 평균이 $\mu_i = \mu(\boldsymbol{x}_i)$인 포아송 확률변수라고 한다면 로그가능도함수는 다음과 같다.

$$\begin{aligned} l &= \log \prod_{i=1}^{n} f(y_i|\mu_i) \\ &= \sum_{i=1}^{n} [y_i \log \mu_i - \mu_i - \log y_i!] \end{aligned} \tag{6.3}$$

위에서 볼수 있듯이 충분통계량 y_i에 대응하는 모수에 대한 항은 $\log \mu_i$ 로서 이는 로그함수가 자연연결함수임을 나타낸다.

회귀계수 $\boldsymbol{\beta}$의 추정은 로지스틱 회귀와 같이 최대가능도추정법(maximum likelihood estimation)으로 구한다. 포아송 회귀에서도 최대가능도추정량은 직접 계산으로 구할 수 없기 때문에 수치적인 방법을 이용하여 구한다.

또한 회귀계수에 대한 검정 $H_0 : \beta_i = 0$은 대표본이론에 의거한 정규근사를 이용한다. 즉 유의수준 α에서 t-통계량 $t = \hat{\beta}_i / se(\hat{\beta}_i)$의 절대값 $|t|$가 z_α보다 크면 귀무가설을 기각한다.

6.4. 편차

포아송 회귀분석에서 편차(deviance) D를 구해보기 위하여 포화모형을 생각해보자.

각 관측값의 평균 μ_i를 자신의 관측값 y_i로 추정하는 것이 포화모형이다. 따라서 포화모형의 로그가능도함수는 다음과 같이 주어지고

$$l_{saturated} = \sum_{i=1}^{n} [y_i \log y_i - y_i - \log y_i!]$$

식 6.3 으로 주어진 포아송 회귀분석의 로그가능도함수를 빼주면 D를 얻을 수 있다.

$$\begin{aligned} D &= 2[l_{saturated}(\hat{\boldsymbol{\mu}}|y) - l_{regession}(\hat{\boldsymbol{\mu}}|y)] \\ &= 2\sum_{i=1}^{n}[y_i \log y_i - y_i - \log y_i!] - \sum_{i=1}^{n}[y_i \log \hat{\mu}_i - \hat{\mu}_i - \log y_i!] \\ &= 2\sum_{i=1}^{n}[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)] \end{aligned}$$

또한 모형의 적합성을 측정하는 양으로서 χ^2-통계량을 사용할 수 있다.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

로지스틱 회귀에서와 비슷하게 포아송 회귀분석에서도 과포화(overdispersion)가 나타날 수 있다. 즉, 포아송 모형의 가정은 평균과 분산이 같은 것인데 ($\mu_i = E(y_i) = Var(y_i)$) 이러한 가정은 실제 자료 분석에서 많은 경우에 만족하지 않을 수 있으며 과포화가 나타난다

$$E(y_i) = \mu_i, \quad \text{but} \quad Var(y_i) > \mu_i$$

이렇게 과포화가 나타나는 경우에는 산포모수(dispersion parameter) ϕ를 추정하여 회귀계수의 표준오차 계산에 반영해주어야 한다. 산포모수는 χ^2 통계량을 통하여 추정할 수 있다.

$$\hat{\phi} = \frac{\chi^2}{n-p}$$

6.5. 발생율 모형

어떤 사건이 일어날 횟수는 집단이나 시간의 크기(size)에 의존할 수 있다. 예를 들어 각 도시의 1년 범죄 발생 횟수는 그 도시의 인구수나 크기에 비례하게 된다.

이러한 모형은 이항분포를 이용하여 분석할 수 도 있지만 사건의 발생확률이 매우 작고 집단의 크기가 크면 포아송 근사를 통한 분석도 가능하다. 또한 어떤 경우에는 집단의 크기에 대한 정보가 부족할 수 있다.

이러한 비율에 대한 회귀모형을 발생율 모형(rate models) 로 부르며 식으로 나타내면 아래와 같고

$$\log \frac{\text{발생횟수}}{\text{집단의 크기}} = \boldsymbol{x}^t\boldsymbol{\beta}$$

이는 다시 발생횟수에 대한 포아송 회귀모형의 형태로 나타내면 다음과 같이 쓸 수 있다.

$$\log \text{ 발생횟수} = (1)(\log \text{집단의 크기}) + \boldsymbol{x}^t\boldsymbol{\beta}$$

따라서 발생횟수에 대한 포아송 회귀분석을 적합할 때 집단의 크기를 안다면 그 log 변환값을 회귀식에 포함하여 적합할 수 있다. 위의 식에서 알 수 있듯이 크기의 log 변환변수는 회귀계수를 강제로 1로 놓는 제약을 둘 수 있다. 이러한 변수를 오프셋 변수(offset variable)이라고 한다.

6.6. 음이항 분포

베르누이 독립시행에서 k번째의 성공까지의 시행회수 z는 음이항 분포(negative bionomial)을 따른다. 음이항분포는 포아송 분포에서 모수가 감마를 따를 때 근사분포로 사용될 수 있다.

$$P(z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}, \quad z = k, k+1, ... \tag{6.4}$$

위의 분포에서 확률 변수와 모수를 다시 아래와 같이 정의하면

$$y = z - k, \quad p = \frac{1}{1+\alpha}$$

y의 확률분포는 다음과 같고

$$P(y) = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1+\alpha)^{y+k}}, \quad y = 0, 1, 2, ...$$

따라서 y의 평균과 분산은 다음과 같이 주어진다.

$$E(y) = \mu = k\alpha, \quad Var(y) = k\alpha + k\alpha^2 = \mu + \mu^2/k$$

또한 로그가능도함수는 다음과 같이 주어지고

$$l = \sum_{i=1}^{n} \left(y_i \log \frac{\alpha}{1+\alpha} - k \log(1+\alpha) + \sum_{j=0}^{y_i - 1} \log(j+k) - \log y_i! \right)$$

연결함수는 다음과 같다.

$$\log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{\mu+k} = \eta = \boldsymbol{x}^t \boldsymbol{\beta}$$

보통의 경우 k는 고정된 상수로 생각할 수도 있고 또는 모수로 보고 추정할 수 도 있다.

6.7. 영과잉모형

어떤 사건의 발생횟수에 대한 자료를 수집할 때 0이 비정상적으로 많이 나타나는 경우가 있다.

만약 발생횟수의 분포를 포아송분포(식 6.1)로 가정하면 0이 관측될 확률은 크지 않다.

$$P(y = 0) = e^{-\mu}$$

자료에서 0의 발생 빈도가 비정상적으로 많은 자료를 영과잉자료(zero inflated data)라고 하며 이러한 자료에 포아송 분포를 그대로 적용하면 회귀 계수의 추정량에 편이(bias)가 발생할 수 있으며 과포화(overdispersion)가 발생하는 등 여러 가지 문제가 생긴다.

발생횟수에 0이 많은 이유는 매우 다양하다. 0이 많이 발생하는 대표적인 이유를 살펴보자.

- 외부 요인에 의하여 사건의 발생이 제약을 받는 경우
- 발생은 했는데 관측이 안된 경우
- 원래 0이 많은 경우

이렇게 0 과잉 자료를 분석할 수 있는 대표적인 모형은 영과잉 포아송 모형(zero inflated poission model; ZIP)이다

확률변수 y_i를 사건의 발생 회수라고 하면 ZIP 모형에서 0이 관측될 확률을 다음과 같이 나타낼 수 있다.

$$P(y=0) = P(\text{ False zeros }) + [1 - P(\text{ False zeros })]P(\text{ count process gives a zero })$$

즉 0이 관측될 확률은 잘못된 0이 관찰 될 확률과 원래 확률 과정에서 0이 관찰 될 확률의 조합(mixture)으로 나타난다. 이제 i번째 관측에서 잘못된 0이 관찰 될 확률을 π_i라 하면

$$P(y_i=0) = \pi_i + (1-\pi_i)P(\text{ count process gives a zero })$$

더 나아가 확률 과정이 평균이 μ_i인 포아송 분포를 따른다고 가정하고

$$\begin{aligned} P(y_i=0) &= \pi_i + (1-\pi_i)P(y_i=0|\mu_i) = \pi_i + (1-\pi_i)e^{-\mu_i} \\ P(y_i=k) &= P(y_i=k|\mu_i) = (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_k}}{y_k!}, \quad k=1,2,\cdots \end{aligned}$$

위의 분포에서 y의 평균과 분산을 구해보면 다음과 같이 주어진다.

$$\begin{aligned} E(y_i) &= (1-\pi_i)\mu_i \\ Var(y_i) &= (1-\pi_i)\mu_i + (1-\pi_i)\pi_i\mu_i^2 \end{aligned}$$

위의 식에서 볼 수 있듯이 영과잉 포아송 모형은 과포화들 보인다

$$Var(y_i) > E(y_i)$$

영과잉 포아송 모형에 대한 회귀분석은 다음 두 모형을 동시에 고려하는 모형이다.

- 잘못된 0이 관측될 확률 π_i에 대한 로지스틱 회귀모형
- 발생회수에 대한 포아송 회귀모형

$$\begin{aligned} \log\frac{\pi_i}{1-\pi_i} &= \boldsymbol{x}_b^t\boldsymbol{\beta}_b \\ \log\mu_i &= \boldsymbol{x}_p^t\boldsymbol{\beta}_p \end{aligned}$$

	Species	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	25.09	346	0.6	0.6	1.84
Bartolome	31	1.24	109	0.6	26.3	572.33
Caldwell	3	0.21	114	2.8	58.7	0.78
Champion	25	0.10	46	1.9	47.4	0.18
Coamano	2	0.05	77	1.9	1.9	903.82
Daphne.Major	18	0.34	119	8.0	8.0	1.84

6.8. 예제

6.8.1. Galapagos 군도의 거북이

Galapagos 군도에 있는 30개의 섬에서 사는 거북이의 개체 수 `Species` 를 반응변수 y로하고 5개의 지리적 변수를 예측변수로 하는 Poisson 회귀식을 적합하려고 한다. 이 예제는 교재 Faraway (2016) 57페이지에 있는 예제이다.

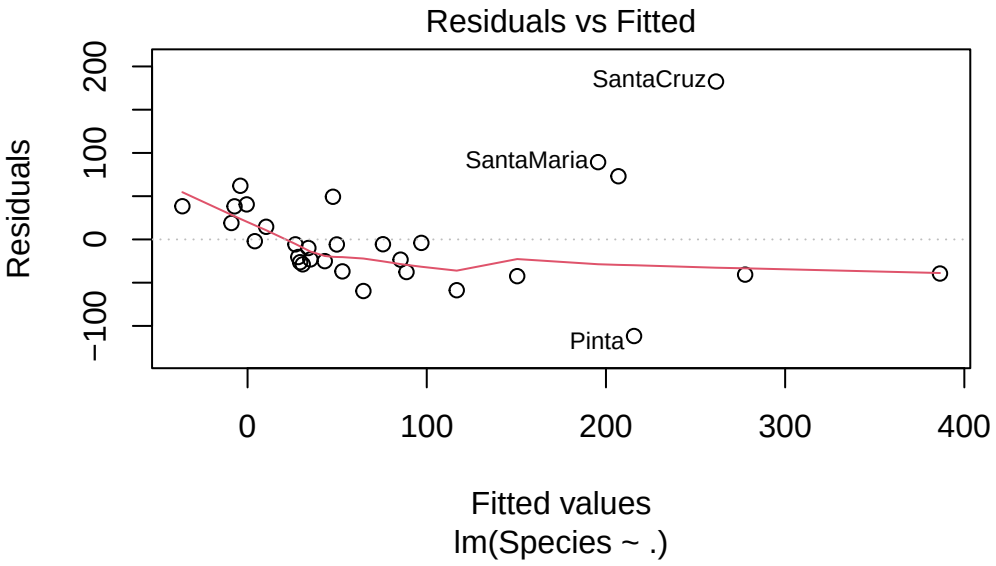
데이터 `gala` 를 구성하는 변수들은 다음과 같다.

- `Species`: the number of plant species found on the island
- `Endemics` : the number of endemic species (아래 분석에서 제외)
- `Area` : the area of the island (km^2)
- `Elevation` : the highest elevation of the island (m)
- `Nearest` : the distance from the nearest island (km)
- `Scruz` : the distance from Santa Cruz island (km)
- `Adjacent` : the area of the adjacent island (square km)

```
gala_2 <- gala[,-2]
head(gala_2)  %>%
  kbl() %>%
  kable_styling( full_width = F)
```

데이터를 일반적인 선형모형과 반응변수에 제곱근 변환을 적용한 결과는 다음과 같다.

```
modl <- lm(Species ~ . , gala_2)
plot(modl, 1)
```



```
summary(modl)
```

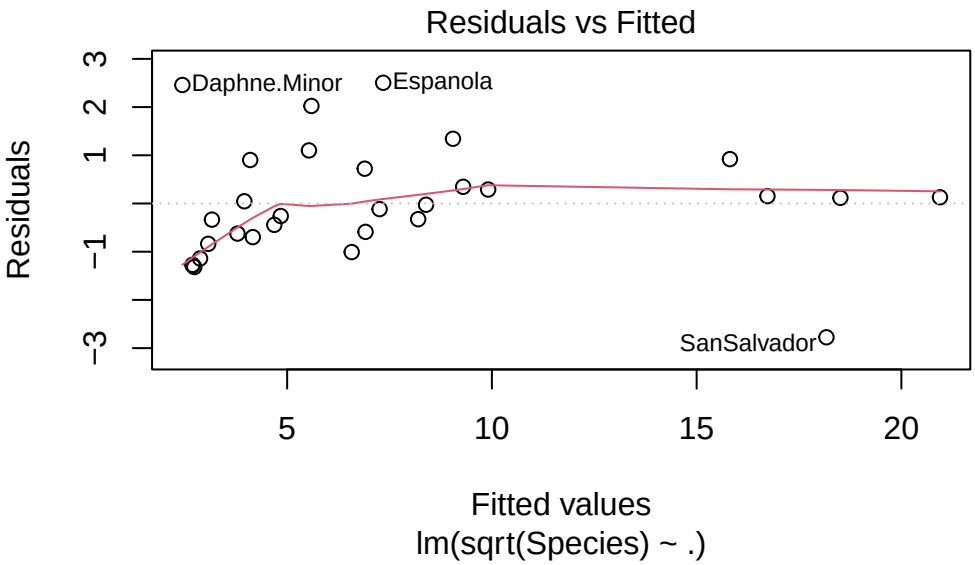
```
Call:
lm(formula = Species ~ ., data = gala_2)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

```
modt <- lm(sqrt(Species) ~ . , gala)
plot(modt, 1)
```




```
summary(modt)
```

```
Call:
lm(formula = sqrt(Species) ~ ., data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-2.77512 -0.67895 -0.07101  0.62771  2.50402

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3705693  0.4253328   5.573 1.14e-05 ***
Endemics     0.2002788  0.0217192   9.221 3.45e-09 ***
Area        -0.0002763  0.0005147  -0.537    0.597
Elevation   -0.0002509  0.0021483  -0.117    0.908
Nearest      0.0198908  0.0226069   0.880    0.388
Scruz       -0.0021423  0.0047791  -0.448    0.658
Adjacent     0.0001255  0.0005361   0.234    0.817
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.307 on 23 degrees of freedom
Multiple R-squared:  0.9537,    Adjusted R-squared:  0.9417
F-statistic: 79.03 on 6 and 23 DF,  p-value: 3.457e-14
```

포아송 회귀모형을 적합한 결과는 다음과 같다.

```
modp <- glm(Species ~ .,family=poisson, gala_2)
summary(modp)
```

```
Call:
glm(formula = Species ~ ., family = poisson, data = gala_2)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
```

cells	ca	doseamt	doserate
478	25	1	0.10
1907	102	1	0.25
2258	149	1	0.50
2329	160	1	1.00
1238	75	1	1.50
1491	100	1	2.00

```
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

6.8.2. 세포의 비정상성

세포(cells)에 감마 방사능을 쏘였을 때 비정상성(ca)를 나타내는 횟수에 대하여 발생율 모형을 적합시켰다. 예측변수는 방사능의 양(doseamt)와 비율(doserate)이다. 여기서 세포의 수(cells)를 오프셋 변수(offset variable)로 사용한다. 이 예제는 교재 Faraway (2016) 61페이지에 있는 예제이다.

- `cells` : Number of cells in hundreds
- `ca` : Number of chromosomal abnormalities
- `doseamt` : amount of dose in Grays
- `doserate` : rate of dose in Grays/hour

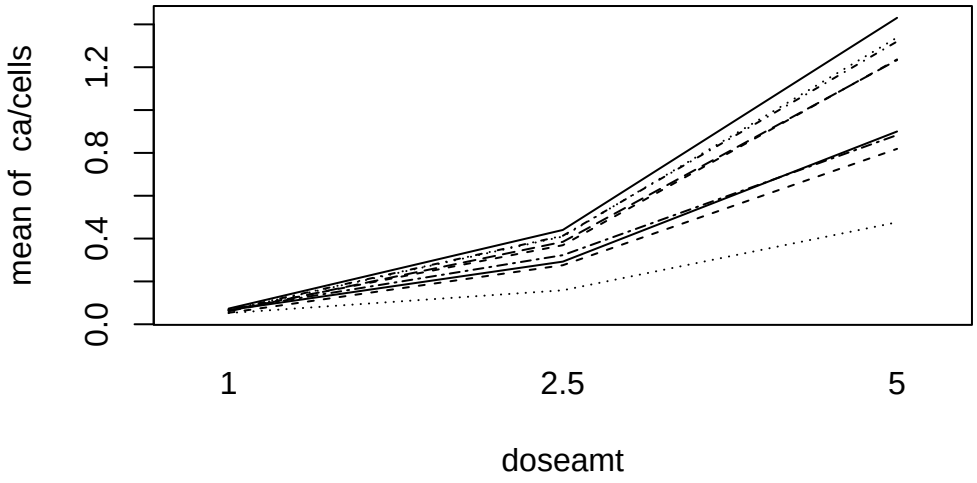
```
head(dicentric) %>%
  kbl() %>%
  kable_styling( full_width = F)
```

다음 표는 방사능의 양(doseamt)와 비율(doserate)의 조합에 따라 비정상 세포의 비율을 나타낸다.

```
round(xtabs(ca/cells ~ doseamt+doserate, dicentric),2)
```

```
       doserate
doseamt  0.1 0.25  0.5    1  1.5    2  2.5    3    4
    1   0.05 0.05 0.07 0.07 0.06 0.07 0.07 0.07 0.07
    2.5 0.16 0.28 0.29 0.32 0.38 0.41 0.41 0.37 0.44
    5   0.48 0.82 0.90 0.88 1.23 1.32 1.34 1.24 1.43
```

```
with(dicentric,interaction.plot(doseamt, doserate, ca/cells, legend= FALSE))
```



먼저 일반적인 선형모형을 적용하고 잔차분석을 수행하자.

```
lmod <- lm(ca/cells ~ log(doserate)*factor(doseamt), dicentric)
summary(lmod)
```

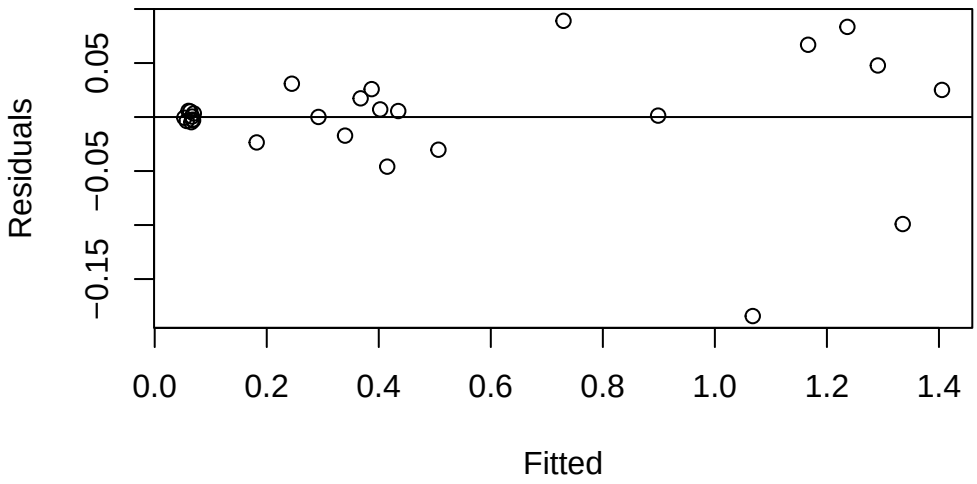
```
Call:
lm(formula = ca/cells ~ log(doserate) * factor(doseamt), data = dicentric)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184275 -0.004212  0.001314  0.021208  0.089076

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.063489   0.019528   3.251  0.00382 **
log(doserate)                     0.004573   0.016692   0.274  0.78680
factor(doseamt)2.5                0.276315   0.027616  10.005 1.92e-09 ***
factor(doseamt)5                  1.004119   0.027616  36.359  < 2e-16 ***
log(doserate):factor(doseamt)2.5 0.063933   0.023606   2.708  0.01317 *
log(doserate):factor(doseamt)5   0.239129   0.023606  10.130 1.54e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05858 on 21 degrees of freedom
Multiple R-squared:  0.9874,    Adjusted R-squared:  0.9844
F-statistic:   330 on 5 and 21 DF,  p-value: < 2.2e-16
```

```
plot(residuals(lmod) ~ fitted(lmod),xlab="Fitted",ylab="Residuals")
abline(h=0)
```



위의 잔차분석에서는 등분산성이 의심되는 결과가 나타난다.

이제 doseamt 를 범주형 자료로 만들고 세포의 갯수(cells) 를 다음과 같이 log 변환하여 독립변수에 포함시키자. 이제 반응변수는 비정상 세포의 비율이 아니라 비정상 세포의 갯수이다.

```
dicentric_1 <- dicentric
dicentric_1$dosef <- factor(dicentric_1$doseamt)
pmod <- glm(ca ~ log(cells)+log(doserate)*dosef,  family=poisson, dicentric_1)
summary(pmod)
```

```
Call:
glm(formula = ca ~ log(cells) + log(doserate) * dosef, family = poisson,
    data = dicentric_1)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.76534    0.38116  -7.255 4.02e-13 ***
log(cells)             1.00252    0.05137  19.517  < 2e-16 ***
log(doserate)          0.07200    0.03547   2.030 0.042403 *
dosef2.5               1.62984    0.10273  15.866  < 2e-16 ***
dosef5                 2.76673    0.12287  22.517  < 2e-16 ***
log(doserate):dosef2.5 0.16111    0.04837   3.331 0.000866 ***
log(doserate):dosef5   0.19316    0.04299   4.493 7.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 916.127  on 26  degrees of freedom
Residual deviance:  21.748  on 20  degrees of freedom
AIC: 211.15

Number of Fisher Scoring iterations: 4
```

이제 다음과 같은 관계에 따라서 세포의 갯수(cells)를 오프셋 변수로 다시 적합시켜보자.

$$\log(\text{ca}/\text{cell}) = \boldsymbol{x}^t\boldsymbol{\beta} \Leftarrow \log(\text{ca}) = \log(\text{cell}) + \boldsymbol{x}^t\boldsymbol{\beta}$$

```
rmod <- glm(ca ~ offset(log(cells))+log(doserate)*dosef,  family= poisson, dicentric_1)
summary(rmod)
```

```
Call:
glm(formula = ca ~ offset(log(cells)) + log(doserate) * dosef,
    family = poisson, data = dicentric_1)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.74671    0.03426 -80.165  < 2e-16 ***
log(doserate)          0.07178    0.03518   2.041 0.041299 *
dosef2.5               1.62542    0.04946  32.863  < 2e-16 ***
dosef5                 2.76109    0.04349  63.491  < 2e-16 ***
log(doserate):dosef2.5 0.16122    0.04830   3.338 0.000844 ***
log(doserate):dosef5   0.19350    0.04243   4.561  5.1e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:   21.75  on 21  degrees of freedom
AIC: 209.16

Number of Fisher Scoring iterations: 4
```

6.8.3. 국립공원 방문자

국립공원에서 일하는 야생동물 연구원은 공원방문자들이 얼마나 많은 수의 고기(fish)를 잡는지 알고 싶어한다.공원방문자는 공원을 떠날 때 다음과 같은 설문들에 대하여 답하였다. 총 250명의 방문자(group)가 설문에 응답하였다. 이 예제는 교재 Faraway (2016) 94페이지에 있는 예제이다.

- 며칠동안 공원에 머물렀는가?
- 같이 공원에 온 사람들은 총 몇 명인가? (persons)
- 같이 공원에 온 사람들 중에 어린이는 몇 명인가? (child)
- 공원 방문중에 낚시를 하였는가?
- 낚시를 하였다면 고기를 몇 마리 잡았는가? (count)
- 공원을 방문할 때 캠핑카를 가지고 왔는가? (camper)

공원에 방문한 사람들 중에 낚시를 하지 않은 사람들이 많기 때문에 많은 수의 그룹이 잡은 고기의 수가 0이다.

위의 자료를 영과잉모형으로 분석하기 위하여 다음과 같이 자료를 부르고 히스토그램을 그려보자.

nofish	livebait	camper	persons	child	xb	zg	count
1	0	0	1	0	-0.8963146	3.0504048	0
0	1	1	1	0	-0.5583450	1.7461489	0
0	1	0	1	0	-0.4017310	0.2799389	0
0	1	1	2	1	-0.9562981	-0.6015257	0
0	1	0	1	0	0.4368910	0.5277091	1
0	1	1	4	2	1.3944855	-0.7075348	0

```
zinb<-read.csv(here::here("data/fish.csv"))
head(zinb)  %>%
  kbl() %>%
  kable_styling( full_width = F)
```

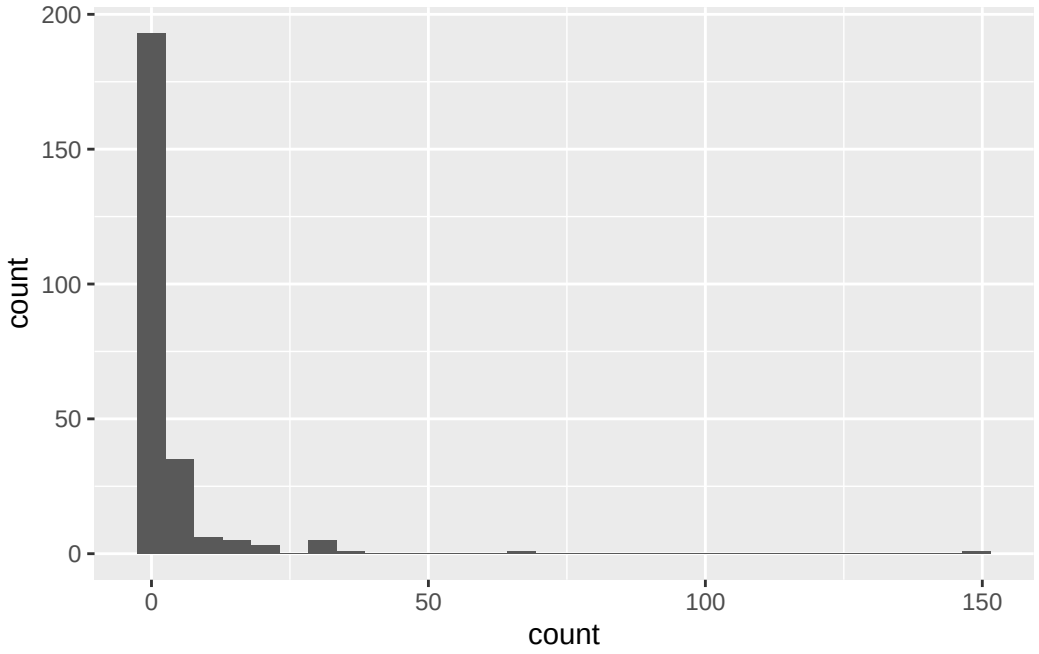
```
zinb_1 <- within(zinb, {
    nofish <- factor(nofish)
    livebait <- factor(livebait)
    camper <- factor(camper)
})

summary(zinb_1)
```

```
 nofish  livebait camper     persons          child             xb
 0:176   0: 34    0:103   Min.   :1.000   Min.   :0.000   Min.   :-3.275050
 1: 74   1:216    1:147   1st Qu.:2.000   1st Qu.:0.000   1st Qu.: 0.008267
                          Median :2.000   Median :0.000   Median : 0.954550
                          Mean   :2.528   Mean   :0.684   Mean   : 0.973796
                          3rd Qu.:4.000   3rd Qu.:1.000   3rd Qu.: 1.963855
                          Max.   :4.000   Max.   :3.000   Max.   : 5.352674
       zg              count
 Min.   :-5.6259   Min.   :  0.000
 1st Qu.:-1.2527   1st Qu.:  0.000
 Median : 0.6051   Median :  0.000
 Mean   : 0.2523   Mean   :  3.296
 3rd Qu.: 1.9932   3rd Qu.:  2.000
 Max.   : 4.2632   Max.   :149.000
```

```
ggplot(zinb_1, aes(count)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

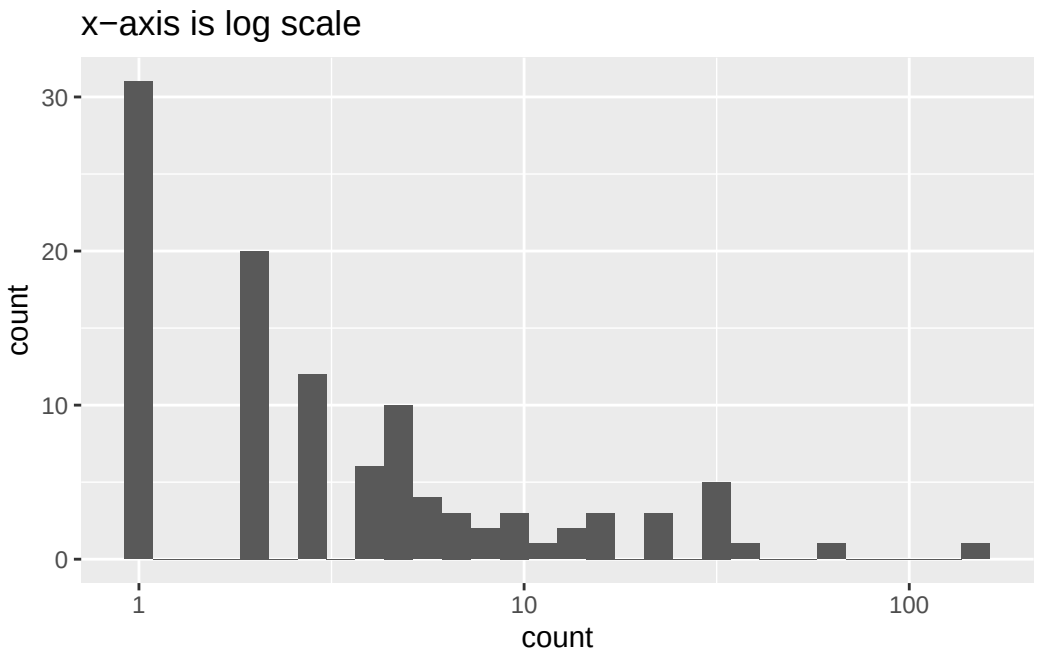



```
ggplot(zinb, aes(count)) + geom_histogram() + scale_x_log10() +ggtitle("x-axis is log scale")
```

```
Warning: Transformation introduced infinite values in continuous x-axis
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 142 rows containing non-finite values (`stat_bin()`).
```



다음은 영과잉 포아송모형을 적합한 결과이다. 0에 대한 로지스틱 회귀에서 사용하는 독립변수는 persons이다.

```
zip1 <- pscl::zeroinfl(count ~ child + camper | persons, data = zinb_1)
summary(zip1)
```

```
Call:
pscl::zeroinfl(formula = count ~ child + camper | persons, data = zinb_1)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.2369 -0.7540 -0.6080 -0.1921 24.0847

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.59789    0.08554  18.680   <2e-16 ***
child       -1.04284    0.09999 -10.430   <2e-16 ***
camper1      0.83402    0.09363   8.908   <2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2974     0.3739   3.470 0.000520 ***
persons      -0.5643     0.1630  -3.463 0.000534 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 10
Log-likelihood: -1032 on 5 Df
```

다음은 일반적인 포아송모형을 적합한 결과이다.

```
pois1 <- glm(count ~ child + camper, family = poisson, data = zinb_1)
summary(pois1)
```

```
Call:
glm(formula = count ~ child + camper, family = poisson, data = zinb_1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.91026    0.08119   11.21   <2e-16 ***
child       -1.23476    0.08029  -15.38   <2e-16 ***
camper1      1.05267    0.08871   11.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 2958.4  on 249  degrees of freedom
Residual deviance: 2380.1  on 247  degrees of freedom
AIC: 2723.2

Number of Fisher Scoring iterations: 6
```

영과잉 모형과 일반적인 포아송 모형의 적합도를 비교하는 가설검정은 Vuong test 를 이용하여 다음과 같이 수행할 수 있다. 아래에서 p-값이 매우 작은 것은 영과잉 모형이 더 적절함을 나타낸다.

```
pscl::vuong(pois1, zip1)
```

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                   -3.574259 model2 > model1 0.00017561
AIC-corrected         -3.552397 model2 > model1 0.00019087
BIC-corrected         -3.513904 model2 > model1 0.00022079
```

References

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Ltd.

———. 2012. *Categorical data analysis*. Vol 792. John Wiley & Sons.

Butler-Laporte, Guillaume, Alexander Lawandi, Ian Schiller, Mandy Yao, Nandini Dendukuri, Emily G McDonald, 와/과 Todd C Lee. 2021. "Comparison of saliva and nasopharyngeal swab nucleic acid amplification testing for detection of SARS-CoV-2: a systematic review and meta-analysis". *JAMA Intern Med* 181 (3): 353–58.

Faraway, Julian J. 2016. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

Weisberg, Sanford. 2014. *Applied Linear Regression*. Fourth. Hoboken NJ: Wiley. http://z.umn.edu/alr4ed.