2018년 2학기 범주형자료분석

통계자료의 유형

양적자료 : 연속형 자료, 키, 몸무게, 소득, 강수량, 자녀의 수 등
질적자료 : 범주형 자료, 명목형(성별, 지역), 순서형(강의평가)

패키지 vcd의 데이터 프레임 Arthritis

```
str(Arthritis)
'data.frame':	84 obs. of  5 variables:
 $ ID       : int  57 46 77 17 36 23 75 39 33 55 ...
 $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...
 $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Age      : int  27 29 30 32 46 58 59 59 63 63 ...
 $ Improved : Ord.factor w/ 3 levels "None"<"Some"<..: 2 1 1 3 3 3 1 3 1 1 ...
head(Arthritis, n=5)
  ID Treatment  Sex Age Improved
1 57   Treated Male  27     Some
2 46   Treated Male  29     None
3 77   Treated Male  30     None
4 17   Treated Male  32   Marked
5 36   Treated Male  46   Marked

# 범주형 변수 : Treatment(Placebo,Treated), Sex(Male, Female), Improved(None, Some, Marked)
# 연속형 변수 : Age
# 설명변수 : Treatment, Sex, Age
# 반응변수 : Improved
```

분할표 작성

table(var1,var2,var3,.....) : N개의 범주형 변수로 N차원 분할표 작성
prop.table(table) : 상대도수 분할표(두 변수의 결합 분포) 작성
prop.table(table,margins) : margins로 정의된 방향으로 조건분포 작성

Improved 에 대한 분할표

```
> with(Arthritis, table(Improved))
Improved
  None   Some Marked
    42     14     28
```

소숫점 자리수 조정

```
> options("digits")
$`digits`
[1] 7
> options("digits"=2)
```

Improved의 상대도수 분포표

```
> my_table1 <- with(Arthritis, table(Improved))
> prop.table(my_table1)
Improved
  None   Some Marked
  0.50   0.17   0.33
```

Treatment와 Improved의 2차원 분할표 및 상대도수 분포표

```
> my_table2 <- with(Arthritis, table(Treatment,Improved))
         Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21

> prop.table(my_table2)
         Improved
Treatment       None       Some     Marked
  Placebo 0.34523810 0.08333333 0.08333333
  Treated 0.15476190 0.08333333 0.25000000
```

2차원 조건분포 분할표 작성

```
# prop.table(table, margin)
# table : 함수 table()로 작성된 분할표
# margin : 조건변수 지정 | margin=1 : 행변수가 조건변수, margin=2 : 열 변수가 조건변수
# 행을 기준으로 다 더하면 1
> prop.table(my_table2, margin=1)
         Improved
Treatment      None      Some    Marked
  Placebo 0.6744186 0.1627907 0.1627907
  Treated 0.3170732 0.1707317 0.5121951
# 열을 기준으로 다 더하면 1
> prop.table(my_table2, margin=2)
         Improved
Treatment      None      Some    Marked
  Placebo 0.6904762 0.5000000 0.2500000
  Treated 0.3095238 0.5000000 0.7500000
```

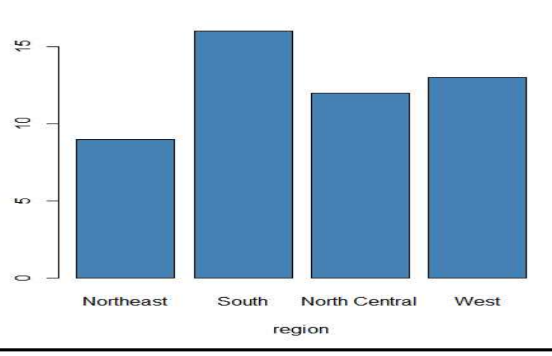
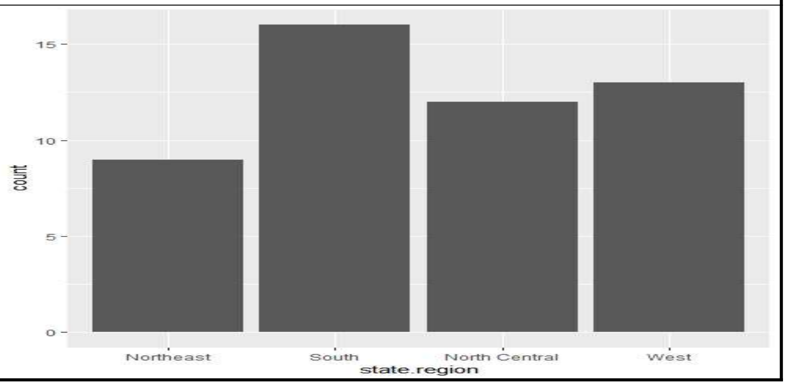
범주형 데이터를 위한 그래프

```
# 분할표 : 자료의 특성을 정확하게 판단하기 어려움
# 자료의 특성 파악을 위해 적절한 그래프 이용이 필수
# 범주형 데이터에 적합한 그래프 : 막대그래프 / 파이그래프 / Mosaic plot(이변량 이상의 경우 적합)

# 막대그래프 작성을 위한 함수
# graphics 패키지 : plot() : 요인을 자료로 입력
#                   barplot() : 도수분포표를 자료 입력
# ggplot2 패키지 : geom_bar() : 요인,도수분포표 모두 사용 가능

# 파이그래프 작성을 위한 함수
# graphics 패키지 : pie() : 도수분포표 자료로 입력
# ggplot2 패키지 : geom_bar() and coord_polar() : 굳이 중요한 그래프는 아님.

# Mosaic plot 작성을 위한 함수
# vcd 패키지에 있는 함수를 사용.
```

예제 데이터 state.region,미국 50개 주를 4개 지역 범주로 구분한 요인

```
> str(state.region)
 Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...

> head(state.region, n=5)
[1] South West  West  South West
Levels: Northeast South North Central West
```

막대그래프 그리기

# base graphics 로 막대그래프 그리기 > plot(state.region, xlab="region", col="steelblue")	# ggplot2 로 막대그래프 그리기 > ggplot(data.frame(state.region), aes(x=state.region)) + geom_bar()
	

만약 그래프를 90도 돌리고 싶다면?

```
> ggplot(data.frame(state.region), aes(x=state.region)) + coord_flip() + geom_bar()
```

도수분포표를 자료로 이용하는 경우

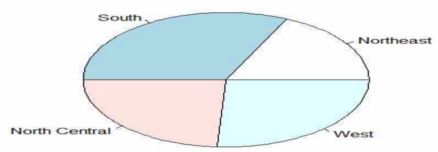
```
> counts <- table(state.region)
> counts
state.region
     Northeast         South    North Central         West
             9            16               12           13
```

# base graphics 로 막대그래프 그리기	# ggplot2 로 막대그래프 그리기
> barplot(counts, col="steelblue")	# 데이터프레임으로 전환 후 작성. # geom_bar()의 디폴트 stat은 "count" 이다. # stat = "indentity" : 데이터를 있는 그대로 그려라. > df_1 <- as.data.frame(counts) > ggplot(df_1, aes(x=state.region, y=Freq)) + geom_bar(stat="identity", fill="skyblue")

파이그래프

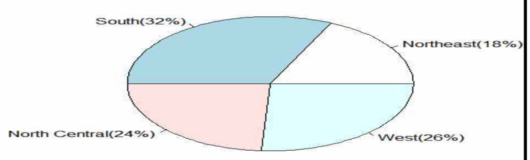
면적으로 빈도수 구분
차이 구분의 정확성 : 길이 vs 면적
좋은 그래프는 아니다.

```
> pie(counts)
```



각 파이 조각에 라벨 추가

```
> pct <- prop.table(counts)*100
> region <- paste0(names(pct), "(",pct,"%)")
> region
[1] "Northeast(18%)"   "South(32%)"   "North Central(24%)"   "West(26%)"
> pie(counts, labels=region)
```



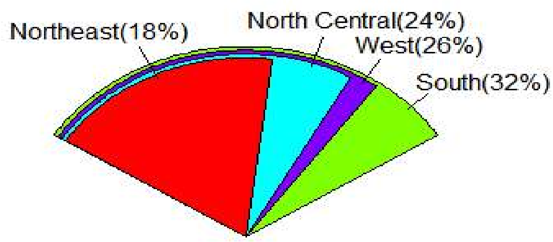
문자열 잇기

```
> # paste : 두 문자열을 이어라.
> x1 <- paste("stat", 1:20)
> head(x1, n=10)
 [1] "stat 1"  "stat 2"  "stat 3"  "stat 4"  "stat 5"  "stat 6"  "stat 7"  "stat 8"  "stat 9"  "stat 10"
> # paste0 : 공백 없이 두 문자를 이어라.
> x2 <- paste0("stat", 1:20)
> head(x2, n=10)
 [1] "stat1"  "stat2"  "stat3"  "stat4"  "stat5"  "stat6"  "stat7"  "stat8"  "stat9"  "stat10"
```

Fan Plot

파이그래프 보다는 가독성이 높다.

```
> library(plotrix)
> fan.plot(counts, labels=region)
```



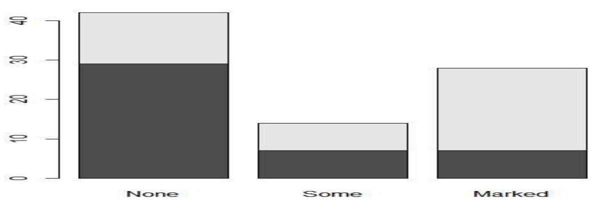
이변량 범주형 자료를 위한 그래프

```
# 막대그래프
# 옆으로 쌓아올린 그래프
# 옆으로 나란한 그래프
# Mosaic 그래프
# 두 개 이상의 범주형 변수 관계 탐색에 유용한 그래프
```

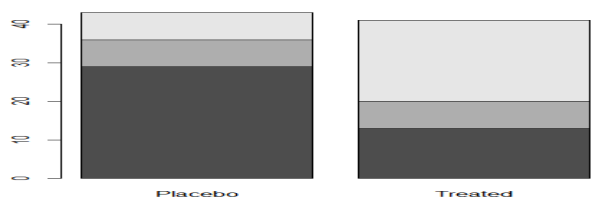
패키지 vcd 의 Arthritis

```
〉 library(vcd)
〉 my_table 〈- with(Arthritis, table(Treatment,Improved))
〉 my_table
          Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21
```

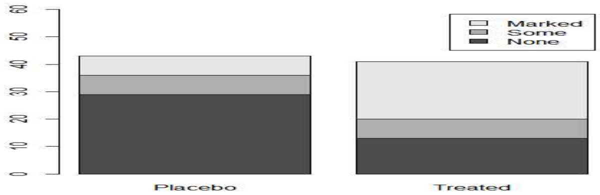
base graphics 로 쌓아 올린 막대 그래프




```
〉 barplot(my_table)
# 함수 barplot() 으로 작성 : 분할표를 자료로 입력
# 설명변수 위주로 쌓아 올리는 것이 더 효과적
```



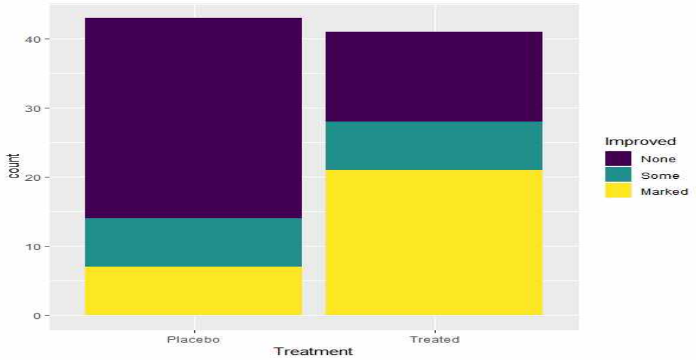

```
〉 barplot(t(my_table))
# 함수 t() 를 사용하면 분할표의 행과 열이 바뀐다.
# 각 조각의 면적은 두 변수의 결합확률에 근거한다.
```



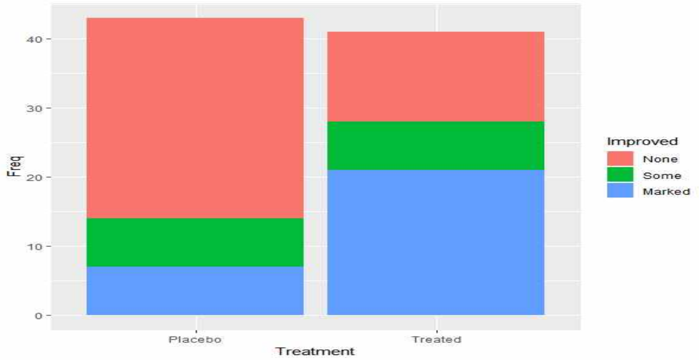

```
〉 barplot(t(my_table), legend.text=TRUE, ylim=c(0,60))
# 쌓아 올려진 조각들의 범례가 필요하다.
# 범례가 겹치는 것을 방지하기 위해 y 의 범위를 늘려준다.
```

ggplot2 의 쌓아 올린 막대 그래프

```
# ggplot2로 작성
# 도형에 색을 채울 때 : fill
# 변수 Improved의 범주별 조각이 쌓아 올려짐.
〉 library(ggplot2)
〉 ggplot(Arthritis, aes(x=Treatment, fill=Improved)) + geom_bar()
# 분할표 사용
〉 ggplot(as.data.frame(my_table), aes(x=Treatment, y=Freq, fill=Improved)) + geom_bar(stat="identity")
```



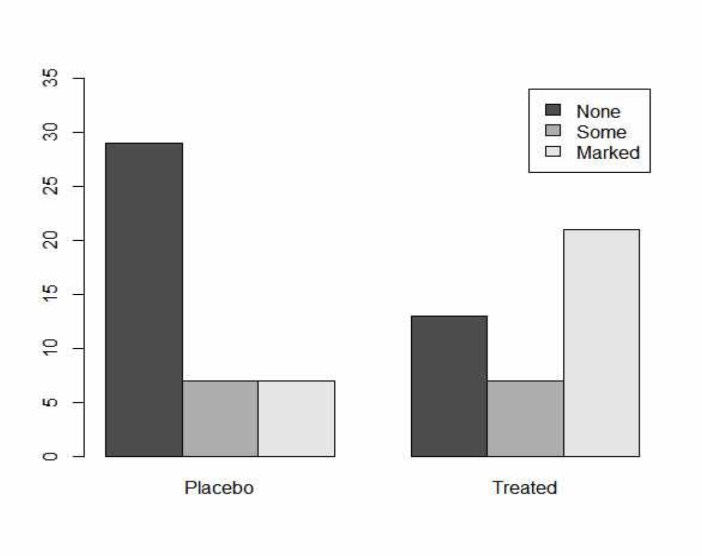
원자료 사용



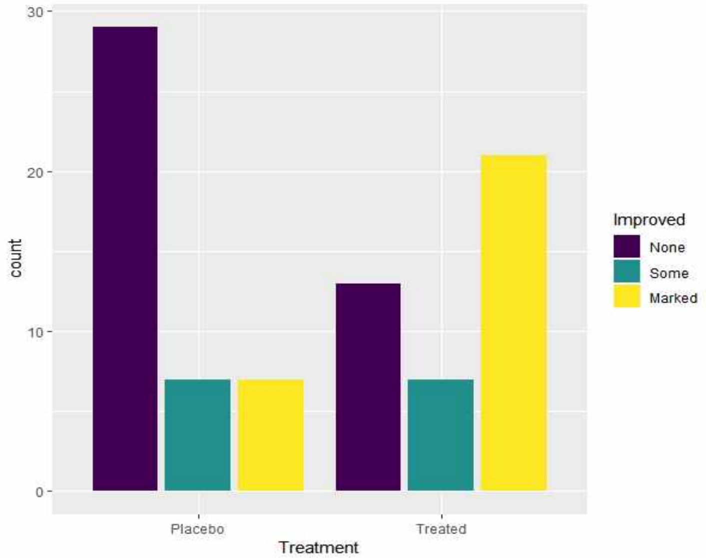
분할표 사용

옆으로 붙여 놓은 막대 그래프

```
# base graphics
〉 barplot(t(my_table), beside=TRUE, legend.text=TRUE, ylim=c(0,35))
# ggplot2
# geom_bar()의 position 디폴트는 : stacked
# dodge : 붙음 / dodge2 : 조금 떨어짐.
〉 pp 〈- ggplot(Arthritis, aes(x=Treatment, fill=Improved))
〉 pp + geom_bar(position="dodge2")
```

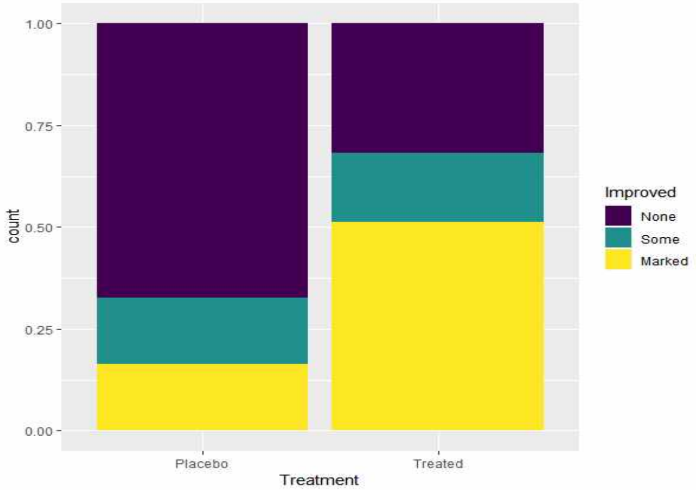


base graphics



ggplot2

geom_bar() 에서 position = "fill" 지정




```
# prop.table(my_table,1)의 결과값을 가지고 그림.
# -〉 행에 대한 조건부 확률을 가지고 그린 것.

〉 pp 〈- ggplot(Arthritis, aes(x=Treatment, fill=Improved))
〉 pp + geom_bar(position="fill")

          Improved
Treatment      None      Some    Marked
  Placebo 0.6744186 0.1627907 0.1627907
  Treated 0.3170732 0.1707317 0.5121951
```

Mosaic Plot (분할표 입력)

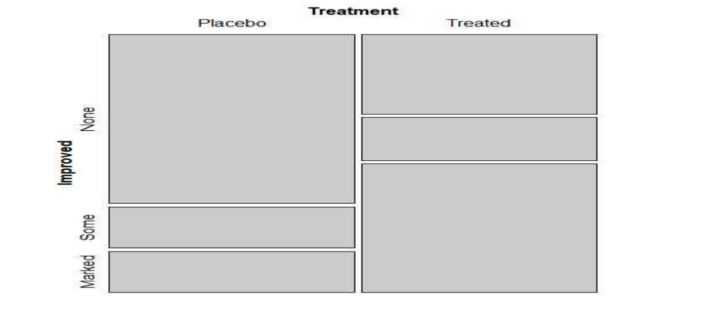
두 개 이상의 범주형 변수 관계 탐색에 유용한 그래프이다.
패키지 vcd의 함수 mosaic() 로 작성한다.

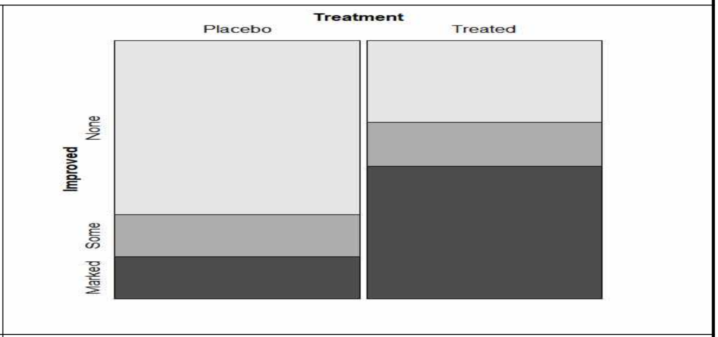
행 변수(Treatment)의 상대도수의 비율로 정사각형을 수직으로 분리(direction= “v”) (direction="h"가 디폴트)
수직을 분리된 두 조각을 행 변수를 조건으로 하는 열 변수의 조건부 확률에 비례하여 수평 방향으로 분리

```
〉 mosaic(my_table, direction="v")
```

Mosaic Plot (원자료 입력)

```
# R 공식으로 변수 선언
# ~변수 + 변수 형태
> mosaic( ~Treatment + Improved, data=Arthritis, direction="v")
# 반응변수-설명변수 형태 : 반응변수의 수준에 따라 조각이 다른 색으로 채워짐
> mosaic(Improved ~ Treatment, data=Arthritis, direction="v")
```





~변수 + 변수 형태	반응변수-설명변수 형태

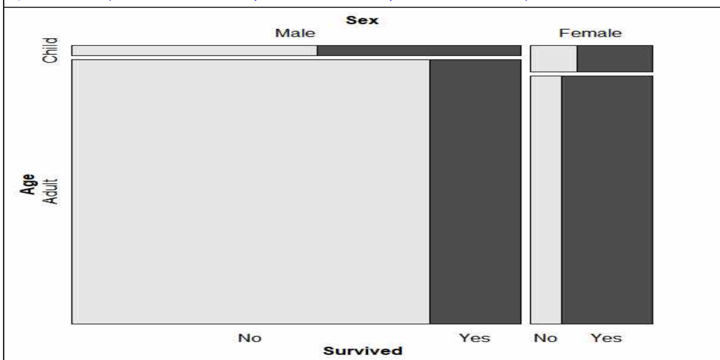
예제 데이터 Titanic

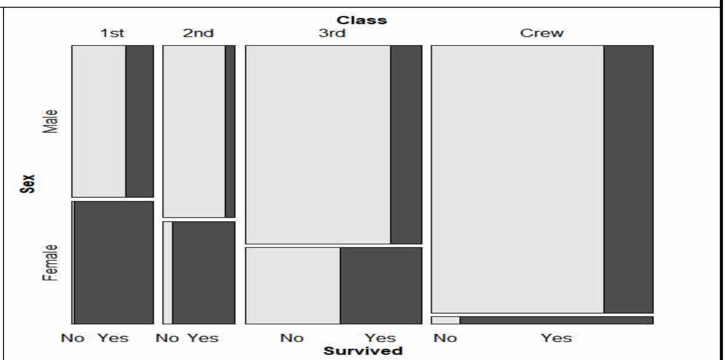
```
> str(Titanic)
 'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
 - attr(*, "dimnames")=List of 4
  ..$ Class    : chr [1:4] "1st" "2nd" "3rd" "Crew"
  ..$ Sex      : chr [1:2] "Male" "Female"
  ..$ Age      : chr [1:2] "Child" "Adult"
  ..$ Survived: chr [1:2] "No" "Yes "
```

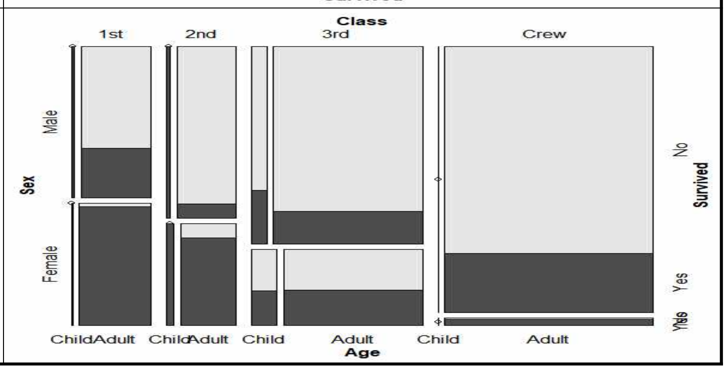
반응변수 : Survived
설명변수 : Class, Sex, Age
생존에 큰 영향을 미친 변수는?

Titanic 의 Mosaic Plot

```
#  "성별 + 나이" 와 "생존" 의 관계
> mosaic(Survived ~ Sex + Age , data=Titanic, direction="v")
#  "좌석등급 + 성별" 과 "생존" 의 관계
> mosaic(Survived ~ Class + Sex , data=Titanic, direction="v")
#  "좌석등급 + 나이" 와 "생존" 의 관계
> mosaic(Survived ~ Class + Age , data=Titanic, direction="v")
#  "모든 변수" 와 "생존" 의 관계
> mosaic(Survived ~ . , data=Titanic, direction="v")
```







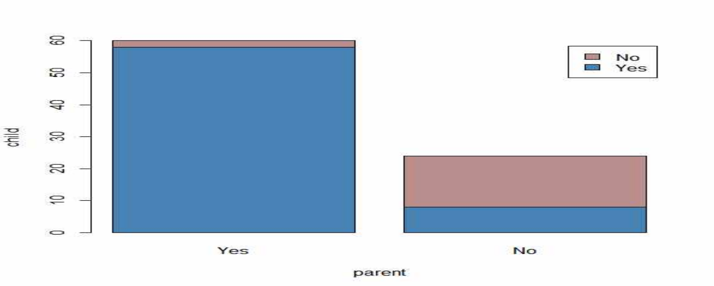
예제 데이터 : 부모와 어린 자녀의 안전벨트 착용 여부에 대한 조사 데이터

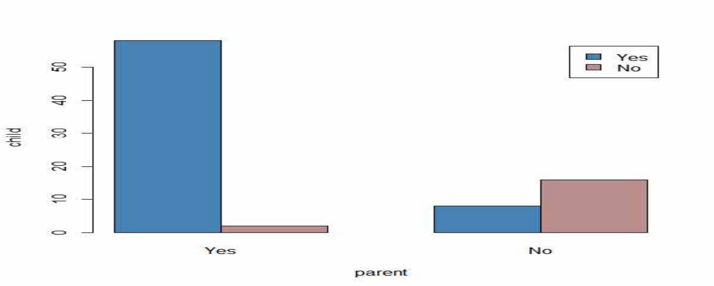
```
# 반응변수 : 아이의 안전벨트 착용 여부
# 설명변수 : 부모의 안전벨트 착용 여부
> belt <- matrix(c(58,8,2,16), nrow=2, ncol=2)
> dimnames(belt) <- list(parent=c("Yes","No"),child=c("Yes","No"))
> belt
```

```
      child
parent Yes No
   Yes  58  2
   No    8 16
```

belt 그래프 (base graphics)

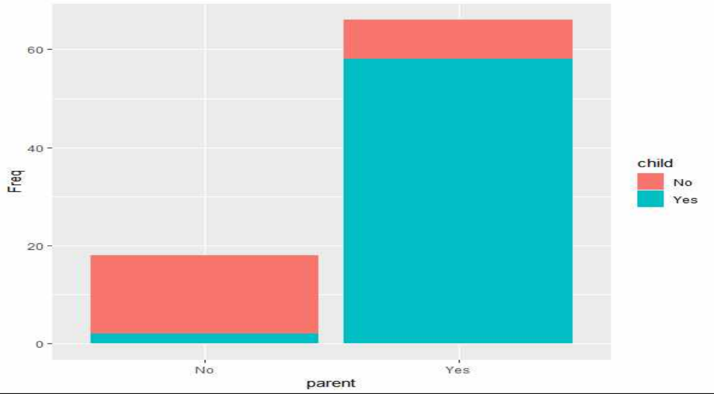
```
# 위로 쌓아올린 그래프
> barplot(t(belt), legend.text=TRUE, xlab="parent", ylab="child", col=c("steelblue","rosybrown"))
# 옆으로 쌓아올린 그래프
> barplot(t(belt), besid=TRUE, legend.text=TRUE, xlab="parent", ylab="child", col=c("steelblue","rosybrown"))
```

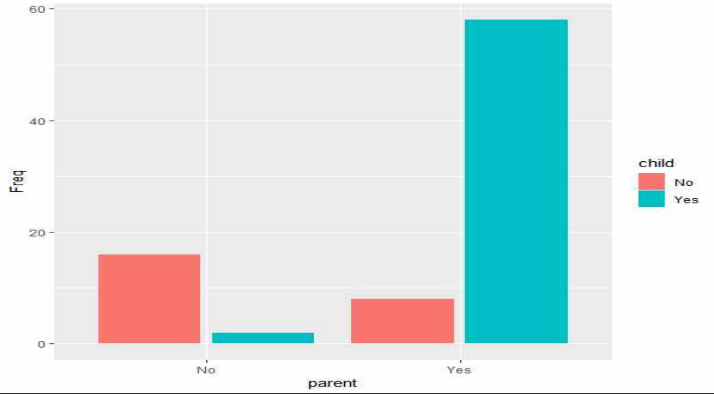


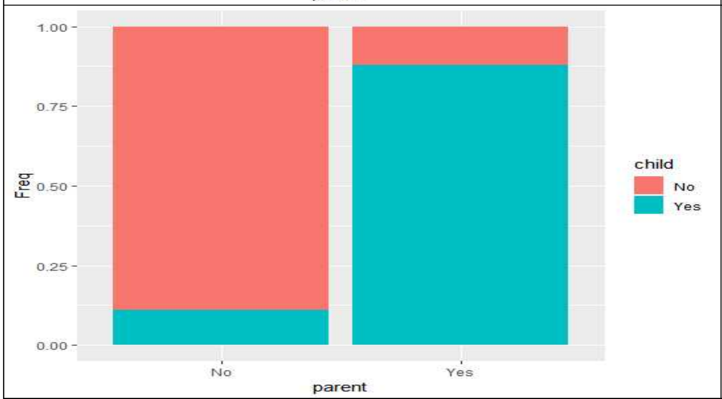


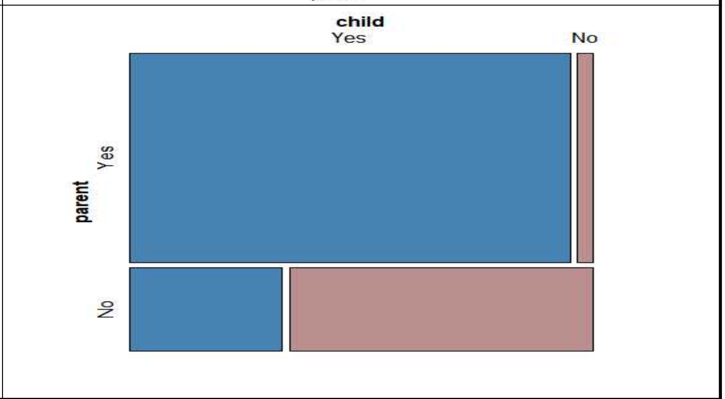
belt 그래프 (ggplot2)

```
# 데이터 프레임 생성
> df_1 <- data.frame(parent=c("Yes","Yes","No","No"), child=c("Yes","No","Yes","No"), Freq=c(58,8,2,16))
> p <- ggplot(df_1, aes(x=parent, y=Freq, fill=child))
# 위로 쌓아올린 막대그래프
> p + geom_bar(stat="identity")
# 옆으로 쌓아올린 막대그래프
> p + geom_bar(stat="identity",position="dodge2")
# 조건부 확률을 이용한 막대그래프
> p + geom_bar(stat="identity",position="fill")
# Mosaic Plot
> mosaic(belt, gp=gpar(fill=c("steelblue","rosybrown")),dirction="v")
```









2차원 $i \times j$ 분할표의 구조 (관찰값 분할표)

		Y				
		1	2	$\cdots$	J	
X	1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
	2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
		n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

n_{ij} : i번째 행, j번째 열의 관찰값 빈도 수

n_{i+} : i번째 행의 빈도 수

n_{+j} : j번째 열의 빈도 수

n : 총 빈도 수

2×2 분할표의 연관성 측도 : Odds ratio

이항변수 : 두 개의 범주를 갖는 범주형 변수
두 이항변수의 연관성 측도 : 오즈비(Odds ratio)
오즈(Odds) : 어떤 사건이 일어날 확률을 일어나지 않을 확률로 나눈 값

$$odds = \frac{P(A)}{1-P(A)}$$

2×2 분할표에서의 오즈비(Odds ratio)

X	Y	
	Success	Failure
1	n_{11}	n_{12}
2	n_{21}	n_{22}

X=1 인 경우, P(Y=Success) = π_1
X=2 인 경우, P(Y=Success) = π_2

X=1 인 경우, Y의 Success odds : $odd1 = \dfrac{\pi_1}{1-\pi_1}$

X=2 인 경우, Y의 Success odds : $odd2 = \dfrac{\pi_2}{1-\pi_2}$

두 odds의 비율인 odds ratio : $\theta = \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$

오즈비의 특성

odds ratio : $\theta = \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$

$0 < \theta < \infty$
두 변수 X, Y가 서로 독립이면, $\pi_1 = \pi_2$ => $\theta = 1$
만일 $\theta > 1$, $\pi_1 > \pi_2$ 이면 X=1 에서의 성공 가능성이 더 높다.
만일 $\theta < 1$, $\pi_1 < \pi_2$ 이면 X=1 에서의 성공 가능성이 더 낮다.

odds raito θ 와 역수 $1/\theta$ 는 두 변수 사이의 같은 정도의 연관성을 보이나, 방향은 반대이다.
$\theta = 0.5$: 첫 행의 odds가 둘째 행 odds의 0.5배 => 둘째 행의 odds 가 첫 행 odds의 1/0.5 = 2배

Odds ratio θ 의 추정량

X	Y	
	Success	Failure
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$\hat{\theta} = \frac{p_1/1-p_1}{p_2/1-p_2} = \frac{n_{11}n_{22}}{n_{12}n_{21}}, \quad p_1 = \frac{n_{11}}{n_{11}+n_{12}}, \quad p_2 = \frac{n_{21}}{n_{21}+n_{22}}$$

Odds ratio 추정량 $\hat{\theta}$ 의 분포 : 오른쪽으로 심하게 치우쳐진 상태 → (0,1)의 구간과 (1,∞)의 구간이 실질적으로 동일함.
효과적인 추론을 위해 추정량의 로그변환이 필요한 상황

로그 오즈비 추정량의 점근적인 분포 : $\log\hat{\theta} \approx N\left(\log\theta, \dfrac{1}{n_{11}} + \dfrac{1}{n_{12}} + \dfrac{1}{n_{21}} + \dfrac{1}{n_{22}}\right)$

$\log\theta$ 에 대한 $100 \times (1-\alpha)\%$ 의 신뢰구간 : $\log\hat{\theta} \pm z_{1-\alpha/2} SE(\log\hat{\theta})$
오즈비 θ 에 대한 신뢰구간 : $\log\theta$ 신뢰구간의 하한과 상한에 지수 역변환을 적용하여 계산
두 이항변수의 독립성 검정($H_0 : \theta = 1,\ H_1 : \theta \neq 1$) 에 사용

연습문제 : Aspirin 복용 여부가 Heart Attack 에 미치는 영향 분석

group	Heart Attack		Total
	Yes	No	
Placebo	189	10845	11034
Aspirin	104	10933	11037

Placebo 그룹의 odds : $189/10845 = 0.0174$
Aspirin 그룹의 odds : $104/10933 = 0.0095$
odds ratio 추정값 : $0.0174/0.0095 = 1.83$
로그 odds ratio 의 95% 신뢰구간 : $0.605 \pm 1.96 \times 0.123 = (0.365, 0.846)$
odds ratio 의 95% 신뢰구간 : $(\exp(0.365), \exp(0.846)) = (1.44, 2.33)$

```
# 패키지 vcd의 함수 oddsratio()
# 기본적인 사용법 : oddsratio(x, log=TRUE)
# x : 2x2 행렬 혹은 table 객체
# log=TRUE : 로그 오즈비 계산(디폴트) / log=FALSE : 오즈비 계산
# 두 이항변수의 독립성 검정 : oddsratio() 로 생성된 객체에 함수 summary() 또는 confint()를 적용

# 자료 입력
> library(vcd)
> aspirin <- matrix(c(189,104,10845,10933), nrow=2, ncol=2, dimnames=list(Group=c("Placebo","Aspirin"),HeartAttack=c("Yes","No")))
> aspirin
         HeartAttack
Group      Yes    No
  Placebo  189 10845
  Aspirin  104 10933
```

$\log\theta$ 의 추론

```
> my_odd1 <- oddsratio(aspirin)
> summary(my_odd1)
z test of coefficients:
                      Estimate  Std. Error  z value    Pr(>|z|)
Placebo:Aspirin/Yes:No 0.60544   0.12284    4.9286   8.282e-07 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

# Odds Ratio의 95% 신뢰구간
> my_odd2 <- oddsratio(aspirin,log=FALSE)
> confint(my_odd2)
                          2.5 %   97.5 %
Placebo:Aspirin/Yes:No 1.440042 2.33078
```

2차원 분할표에 대한 독립성 검정

두 범주형 변수의 독립성 검정
Pearson 카에제곱 검정 (대표본의 경우)
Fisher 의 정확검정 (소표본의 경우)

두 범주형 변수의 분포

결합분포 (Joint distribution)
-> $\pi_{ij} = P(X=i,\ Y=j)$

한계분포 (Marginal distribution)
-> $\pi_{i+} = P(X=i)$, $\pi_{+j} = P(Y=j)$

		Y				
		1	2	$\cdots$	J	
X	1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
	2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_{I+}
		π_{+1}	π_{+2}	$\cdots$	π_{+J}	1

독립성

사건의 독립성 : $P(A \cap B) = P(A) \times P(B)$
확률변수의 독립성 : $P(X=x, Y=y) = P(X=x) \times P(Y=y)$
두 범주형 변수 X와 Y의 독립성 : $\pi_{ij} = \pi_{i+} \times \pi_{+j}\quad i=1,...,I,\ j=1,...,J,$

Pearson 카이제곱 독립성 검정

H_0 : 두 범주형 변수는 서로 독립이다. H_1 : 두 범주형 변수는 서로 독립이 아니다.
$H_0 : \pi_{ij} = \pi_{i+} \times \pi_{+j}$ $H_1 : \pi_{ij} \neq \pi_{i+} \times \pi_{+j}$
관찰 빈도수 : n_{ij}
귀무가설에서의 기대 빈도수 : $\mu_{ij} = n\pi_{ij}$
귀무가설이 사실인 경우 : $n_{ij} - \mu_{ij} \approx 0$
검정통계량 : $\chi^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$ $\hat{\mu}_{ij} = n \times p_{i+} \times p_{+j} = \frac{n_{i+} n_{+j}}{n}$
귀무가설에서 검정통계량의 점근분포 : $\chi^2(df)$, $df = (I-1)(J-1)$
카이제곱 분포를 사용하기 위해서는 대표본이 필수적 → $\mu_{ij} \geq 5$ 의 만족이 필요하다.

R에서의 Pearson 카이제곱 독립성 검정

chisq.test(x, y=NULL, simulate.p.value=FALSE)
x, y : 두 범주형 변수를 나타내는 벡터, 만일 x가 행렬 또는 table 객체이면 y는 무시됨.
simulate.p.value=FALSE : 검정통계량의 근사분포로 카이제곱 분포를 사용하여 p값 계산.
simulate.p.value=TRUE : 모의실험을 통하여 p값 계산. 소규모의 표본에 적합.

```
> aspirin
         HeartAttack
Group       Yes    No
  Placebo  189  10845
  Aspirin  104  10933
```

Yate's continuity correction
2x2 분할표에서만 적용
이산형인 이항분포를 연속형인 카이제곱 분포로 근사할 때의 오류 감소 효과.
표본 수가 너무 작은 경우에는 생략(correct=FALSE).

```
> chisq.test(aspirin)

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

data:  aspirin
X-squared = 24.429, df = 1, p-value = 7.71e-07
```

vcd::Arthritis 의 Treatment 와 Improved 의 독립성 검정

분할표에 의한 카이제곱 검정

```
> library(vcd)
> my_table <- with(Arthritis, table(Treatment,Improved))
> chisq.test(my_table)

        Pearson's Chi-squared test

data:  my_table
X-squared = 13.055, df = 2, p-value = 0.001463
```

두 범주형 변수 입력에 의한 카이제곱 검정

```
> with(Arthritis, chisq.test(Treatment, Improved))

        Pearson's Chi-squared test

data:  Treatment and Improved
X-squared = 13.055, df = 2, p-value = 0.001463
```

Fisher 의 정확검정

Peason 카이제곱 검정은 표본크기가 충분히 큰 경우 적용 가능한 방법
표본크기가 작은 경우 근사분포를 사용하지 않는 방법이 필요

Fisher 의 찻잔

어떤 영국 부인이 milk tea를 만들 때 찻잔에 차를 먼저 붓고 우유를 나중에 부었는지 아니면 우유를 먼저 붓고 차를 나중에 부었는지 맛으로 구분할 수 있다고 주장하였다. 이 주장을 검정하기 위하여 두 가지 방법으로 각각 4잔의 차를 만들고 맛으로 보게하여 다음의 결과를 얻었다.

Guess	Truth		합계
	Milk	tea	
Milk	3	1	n_{1+}
Tea	1	3	n_{2+}
합계	n_{+1}	n_{+2}	n

8잔 중 6잔을 올바르게 구분.

Fisher 의 찻잔

Fisher 의 검정 절차
- 2x2 분할표의 행과 열의 합계는 모두 고정
- n_{11}만 결정되면 나머지 칸 모두 결정
- n_{11}이 가질 수 있는 값은 0,1,2,3,4 -> 각 값을 가질 확률은 초기하 분포로 결정
- n 개의 잔 중 Milk 가 먼저 들어간 n_{+1} 개의 잔을 선택하는 경우에 수에서, Milk guess n_{1+} 중 n_{11} 이 실제 Milk

  따라서 Tea Guess n_{2+} 중 $n_{+1}-n_{11}$ 이 실제 Milk 일 확률은 $P(n_{11})=\frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$

부인의 주장을 검정하기 위한 가설
- H_0 : 맛으로 구분할 수 없다. $(\theta=1)$
- H_1 : 맛으로 구분할 수 있다. $(\theta>1)$

위 가설에 대한 p-값 : 실험 결과 얻어진 n_{11} 의 값 보다 대립가설에서 설정된 방향으로 더 극단적인 값을 취하게 될 확률

초기하분포에서 계산 -> p-값 : $P(n_{11}=3)+P(n_{11}=4)$

R에서 FIsher 정확검정

```
# 초기하분포의 p값 계산-> 함수 dhyper() 사용
# m=4, n=4 의 바구니에서 k=4의 공을 꺼내는 경우, x = 3,4 의 확률을 계산
> dhyper(x=3,m=4,n=4,k=4) + dhyper(x=4,m=4,n=4,k=4)
[1] 0.2428571
```

fisher.test() 이용

```
# fisher.test(x, y=NULL, or=1, alternative="two.sided", conf.int=TRUE, simulate.p.value=FALSE)
# x : 요인 객체 혹은 행렬, table 객체
# y : 요인 객체, x가 행렬이면 무시
# simulate.p.value : 분할표가 2x2 보다 큰 경우, p값을 모의실험을 통해 계산할 것인지 여부

# 나머지 옵션은 2x2 분할표에서만 적용
# or=1 : 귀무가설에서 설정되는 Odds ratio 값
# alternative : 대립가설, 디폴트 값 외에 "less", "greater" 가능
# confint : Odds ratio에 대한 신뢰구간

> TeaTaste <- matrix(c(3,1,1,3), nrow=2, ncol=2, dimnames=list(Guess=c("Milk","Tea"),Truth=c("Milk","Tea")))
> TeaTaste
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3
> fisher.test(TeaTaste, or=1, alternative="greater", conf.int=TRUE, simulate.p.value=FALSE)
          Fisher's Exact Test for Count Data
data:  TeaTaste
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309

# odds ratio의 계산 방식이 앞에서 정의된 것과 다르다. (무시해도 됨.)
# p-값이 0.2429 로 계산되었다.
# 두 변수 Guess 와 Truth 사이의 통계적 양의 연관성을 확립할 수 없다.
# 비록 8 잔 중 6잔을 올바르게 구분하였으나 부인의 주장이 통계적으로는 입증되지 않았다.
```

직업 만족도와 수입의 연관성

데이터

Income	Satisfaction			
	veryD	LittleD	ModerateS	VeryS
〈 15k	1	3	10	6
15 – 25k	2	3	10	7
25 – 40k	1	6	14	12
〉 40k	0	1	9	11

```
> Job <- matrix(c(1,2,1,0,  3,3,6,1,  10,10,14,9,  6,7,12,11), ncol=4,
                    dimnames=list(income=c("<15k","15-25k","25-40k",">40k"), satisfaction=c("VeryD","LiteD","LittleS","veryS")))
> Job
        satisfaction
income   VeryD LiteD LittleS veryS
  <15k       1     3      10     6
  15-25k     2     3      10     7
  25-40k     1     6      14    12
  >40k       0     1       9    11

> chisq.test(Job)
        Pearson's Chi-squared test
data:  Job
X-squared = 5.9655, df = 9, p-value = 0.7434
Warning message:
In chisq.test(Job) : 카이제곱 approximation은 정확하지 않을수도 있습니다.

# 주어진 분할표가 2X2를 초과. odds ratio 의 검정은 불가능
# Pearson 카이제곱 검정과 Fisher의 정확검정으로 독립성여부 확인
# 두 변수의 범주 개수에 비하여 표본 수가 매우 적은 경우 카이제곱 검정에 문제가 발생 할 수 있다.
```

기대빈도수 확인

```
# 분할표의 전체 칸 중 50% 칸의 기대 빈도수가 5 미만 => 카이제곱 분포를 사용하는데 문제가 있음.
> job.t$expected
        satisfaction
income      VeryD    LiteD   LittleS  veryS
  <15k    0.8333333 2.708333  8.958333  7.500
  15-25k 0.9166667 2.979167  9.854167  8.250
  25-40k 1.3750000 4.468750 14.781250 12.375
  >40k    0.8750000 2.843750  9.406250  7.875

# 대안 1 : p-값을 카이제곱 분포가 아닌 모의실험을 통해 계산
# 대안 2 : Fisher의 정확검정 적용
# 대안 3 : 두 범주형 변수의 범주 개수를 축소하여 카이제곱 검정 적용
```

대안1 : 모의실험에 의한 p-값 계산

```
# 모의실험에 의한 것이기 때문에 실행마다 p-값에 약간의 차이가 날 수 있다.
> chisq.test(Job,simulate.p.value=TRUE)
        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)
data:  Job
X-squared = 5.9655, df = NA, p-value = 0.7521
```

대안2 : Fisher의 정확검정 적용

```
> fisher.test(Job, or=1, alternative="two.sided", conf.int=TRUE, simulate.p.value=FALSE)
        Fisher's Exact Test for Count Data
data:  Job
p-value = 0.7827
alternative hypothesis: two.sided
```

대안 3 : 두 범주형 변수의 범주 개수를 축소하여 카이제곱 검정 적용

```
# 변수 Income 범주 2개로 축소
# 〈15k + 15-25k : 〈25k
# 25k-40K + 〉40k : 〉25k

# 변수 Satisfaction 범주 2개로 축소
# VeryD + LittleD : D
# Moderates + VeryS : S

# 패키기 vcdExtra에 있는 함수 collapse.table() 사용
# 4X4 분할표를 2X2 분할표로 축소

〉 library(vcdExtra)
〉 Job 〈- matrix(c(1,2,1,0,  3,3,6,1,  10,10,14,9,  6,7,12,11), ncol=4,
+                dimnames=list(income=c("〈15k","15-25k","25-40k","〉40k"), satisfaction=c("VeryD","LitleD","LittleS","veryS")))
〉 Job.r 〈- collapse.table(as.table(Job), income=c("〈25k","〈25k","〉25k","〉25k"), satisfaction=c("D","D","S","S"))
〉 Job.r
      satisfaction
income  D  S
  〈25k  9 33
  〉25k  8 46

〉 chisq.test(Job.r)
          Pearson's Chi-squared test with Yates' continuity correction
data:  Job.r
X-squared = 0.32791, df = 1, p-value = 0.5669

# 카이제곱 분포의 근사 부정확성 문제는 해결
# 범주의 개수 축소 : 몇 개의 범주를 결합시켜 새로운 범주를 만드는 작업
                    -〉 범주의 특성을 그대로 유지할 수 있도록 하는 것이 중요
```

일반화 선형모형 (Generalized Linear Model)

# 통상적인 회귀모형 -> 반응변수 : 연속형(정규분포 가정) -> 설명변수 : 연속형, 범주형 가능	# 반응변수가 연속형이 아닌 예 -> 이항 변수(성공/실패), 다항 변수(상/중/하) -> Count data(특정 도로 통과 차량 대수)

일반화 선형모형
-> 반응변수 : 연속형 및 범주형 변수 등이 가능
-> 매우 포괄적인 선형모형

통상적인 선형회귀모형의 한계점

모형 : $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$

반응변수 Y의 분포 : 정규분포
-> 정규분포가 아닌 경우의 예 : 특정 도로를 이용하는 자동차 대수(포아송 분포)
특정 실험의 성공/실패 여부(베르누이 분포)
반응변수와 설명변수의 관계 : 선형
-> 비선형 관계의 예 : 새로 출시된 제품의 판매량 추이

GLM 의 3가지 성분

Random Component
Systematic Component
Link Function

$$\underset{\text{Random component}}{Y} \quad \xleftrightarrow{\text{Link Function}} \quad \underset{\text{Systematic Component}}{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}}$$

Random Component

반응변수 Y의 확률분포 규정
-> GLM에서 반응변수 Y의 분포는 Exponential family에 속해야 한다.
Exponential family : 정규분포, 포아송분포, 이항분포, 감마분포 등등

Systematic Component

반응변수에 대한 설명변수의 영향력을 표현
-> 설명변수($x_1, \cdots, x_p$)의 선형결합(Linear Predictor)
-> $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Link Function

Random 성분과 Systematic 성분의 연결
-> 반응변수 Y의 평균 $E(Y) = \mu$가 설명변수의 선형결합 η와 어떻게 연결되어 있는지를 규정하는 함수
-> $g(\mu) = \eta$
반응변수의 분포에 따라 대표적으로 사용되는 link function이 존재
(1) 정규분포 : Identity link, $\mu = \eta$

(2) 포아송분포 : Log link, $\log(\mu) = \eta$

(3) 이항분포 : Logit link, $\log(\frac{\mu}{1-\mu}) = \eta$

이항 반응변수에 대한 선형회귀모형

이항 반응변수 : 두 가지 범주만을 갖는 변수. 일반적으로 1 혹은 0의 값을 부여한다.
이항 반응변수의 분포 : Bernoulli 분포
-> $P(Y=y) = \pi^y (1-\pi)^{1-y}, \quad y = 0, 1$
이항 반응변수의 평균과 분산
-> $E(Y) = \sum_y y \times P(Y=y) = P(Y=1) = \pi$
-> $Var(Y) = E(Y^2) - (E(Y))^2 = \pi(1-\pi)$

이항 반응변수에 대한 선형회귀모형설정 1

Classical Linear Regression Model : GLM with identity link

-> $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{pi} X_{pi} + \epsilon_i \sim N(\mu, \sigma^2)$

-> $Y_i = 0, 1$ -> 오차항의 가정을 만족시킬 수 없음.

Random Component 와 Systematic Component의 범위가 다름

-> $E(Y_i) = P(Y_i = 1) = \pi_i, \quad 0 \le \pi_i \le 1$

-> $E(Y_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{pi} X_{pi}$ 의 범위는 $(-\infty, \infty)$

=> 이항반응변수에 대해서는 Classical Linear Regression model을 적용시킬 수 없다.

예제 데이터 Mroz

```
# carData::mroz : 결혼한 미국 백인 여성의 직업참여 여부 분석
# 반응변수 : lfp(labor-force participation) : yes or no
# 설명변수 : k5 : 5세 이하 자녀 수
             k618 : 6~18세 자녀 수
             age : 부인의 나이
             wc : 부인 대학 교육 여부  (yes or no)
             hc : 남편 대학 교육 여부 (yes or no)
             lwg : 부인 기대 소득의 로그값 (직업이 없는 경우, 다른 변수를 이용한 예측 값)
             inc : 부인 소득을 제외한 가계 소득
> library(carData)
> str(Mroz)
'data.frame':   753 obs. of  8 variables:
 $ lfp : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ k5  : int  1 0 1 0 1 0 0 0 0 0 ...
 $ k618: int  0 2 3 3 2 0 2 0 2 2 ...
 $ age : int  32 30 35 34 31 54 37 54 48 39 ...
 $ wc  : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 2 1 1 1 ...
 $ hc  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ lwg : num  1.2102 0.3285 1.5141 0.0921 1.5243 ...
 $ inc : num  10.9 19.5 12 6.8 20.1 ...
> summary(Mroz)
  lfp          k5              k618            age          wc        hc          lwg               inc
 no :325   Min.   :0.0000   Min.   :0.000   Min.   :30.00   no :541   no :458   Min.   :-2.0541   Min.   :-0.029
 yes:428   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   yes:212   yes:295   1st Qu.: 0.8181   1st Qu.:13.025
           Median :0.0000   Median :1.000   Median :43.00                       Median : 1.0684   Median :17.700
           Mean   :0.2377   Mean   :1.353   Mean   :42.54                       Mean   : 1.0971   Mean   :20.129
           3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00                       3rd Qu.: 1.3997   3rd Qu.:24.466
           Max.   :3.0000   Max.   :8.000   Max.   :60.00                       Max.   : 3.2189   Max.   :96.000
```

예제 데이터 Mroz의 선형회귀모형 추정 및 검정

데이터 확인

```
> mroz <- mutate(Mroz, lfp=as.numeric(lfp)-1)
> head(mroz, n=3)
  lfp k5 k618 age wc hc       lwg   inc
1   1  1    0  32 no no 1.2101647 10.91
2   1  0    2  30 no no 0.3285041 19.50
3   1  1    3  35 no no 1.5141279 12.04
```

Mroz 데이터를 통한 선형회귀모형 적용
먼저 k5만 설명변수로 사용
추정대상은 lfp가 yes일 확률
변수 lfp는 factor with 2 levels(no,yes)
함수 lm()에서는 반응변수는 반드시 숫자형
변수 lfp를 숫자형으로 변환 no->0, yes->1

회귀모형 : $E(Y) = p(Y=1) = \beta_0 + \beta_1 X_1$

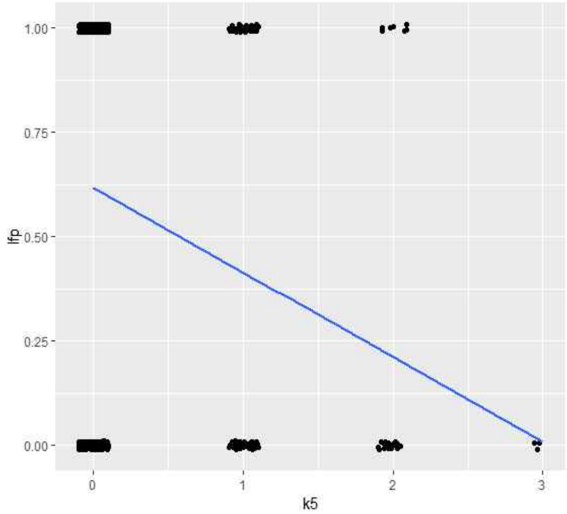
```
> fit <- lm(lfp~k5, data=mroz)
> summary(fit)
Call:
lm(formula = lfp ~ k5, data = mroz)
Residuals:
    Min      1Q  Median      3Q     Max
-0.6165 -0.6165  0.3835  0.3835  0.7879
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.61646    0.01939  31.792  < 2e-16 ***
k5          -0.20219    0.03372  -5.996 3.14e-09 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
Residual standard error: 0.4845 on 751 degrees of freedom
Multiple R-squared:  0.04569,  Adjusted R-squared:  0.04442
F-statistic: 35.96 on 1 and 751 DF,  p-value: 3.136e-09
```

결정계수가 0.04 로 매우 낮다.
회귀계수는 유의하다.

추정된 회귀직선 및 반응변수의 관찰값

관찰값의 개수는 753개이다. 점이 7개만 찍힌 것이 아니라 겹쳐있는 것이다.
겹쳐있는 점을 흐트리기 : geom_jitter(), 점의 위치에 random noise를 추가

```
> ggplot(data=mroz, aes(x=k5, y=lfp)) +
    geom_jitter(height=0.01, width=0.1) + geom_smooth(method="lm", se=FALSE)
```



추정된 회귀모형의 문제점

5세 이하 자녀의 수(k5)가 증가함에 따라 부인이 직업을 가질 확률을 감소
k5=4 인 경우, 확률값이 음수로 추정된다. -> 회귀모형의 적합성에 중대한 문제

k5=4 인 경우의 적합값 예측 (95% 예측 구간 포함)

```
> predict(fit, newdata=data.frame(k5=c(4)), interval="confidence", level=0.95)
         fit        lwr        upr
1 -0.1923117 -0.4437612 0.05913788
```

모든 설명변수 포함된 회귀모형 적합

```
> fit <- lm(lfp~., data=mroz)
> summary(fit)

Call:
lm(formula = lfp ~ ., data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9268 -0.4632  0.1684  0.3906  0.9602
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.143548   0.127053   9.001  < 2e-16 ***
k5          -0.294836   0.035903  -8.212 9.58e-16 ***
k618        -0.011215   0.013963  -0.803 0.422109
age         -0.012741   0.002538  -5.021 6.45e-07 ***
wcyes        0.163679   0.045828   3.572 0.000378 ***
hcyes        0.018951   0.042533   0.446 0.656044
lwg          0.122740   0.030191   4.065 5.31e-05 ***
inc         -0.006760   0.001571  -4.304 1.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.459 on 745 degrees of freedom
Multiple R-squared:  0.1503,	Adjusted R-squared:  0.1423
F-statistic: 18.83 on 7 and 745 DF,  p-value: < 2.2e-16
```

설정된 회귀모형은 유의적, 그러나 지나치게 낮은 설명력(0.14) 이다.
-> 잘못 설정된 회귀모형의 함수 형태가 원인일 가능성이 높다.

이항 반응변수에 대한 선형회귀모형설정 2

일반화 선형모형(GLM) 적용

-> Random component : 반응변수 Y의 분포
Bernoulli 분포는 Exponential family에 속한다.

-> Systematic component : 설명변수의 선형결합
$\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$

-> Link Function : $E(Y_i)$와 η_i의 관계 설정
$g(\pi_i) = \eta_i$를 설정하는 함수 g 선택. 단, $0 \le g^{-1}(\eta_i) = \pi_i \le 1$ 를 만족

(1) Logit Link : $\log(\frac{\pi}{1-\pi}) = \eta$

(2) Probit Link : $\Phi^{-1}(\pi) = \eta$, Φ^{-1} 는 누적정규분포의 역함수

Link Function 1 : Logit link

성공확률 : $P(Y=1) = \pi$

Odds : $\Omega = \frac{P(Y=1)}{1-P(Y=1)}$, $0 \le \log\Omega \le \infty$

Logit Fuction : $\log\Omega = \log\left[\frac{P(Y=1)}{1-P(Y=1)}\right]$, $-\infty \le \log\Omega \le \infty$

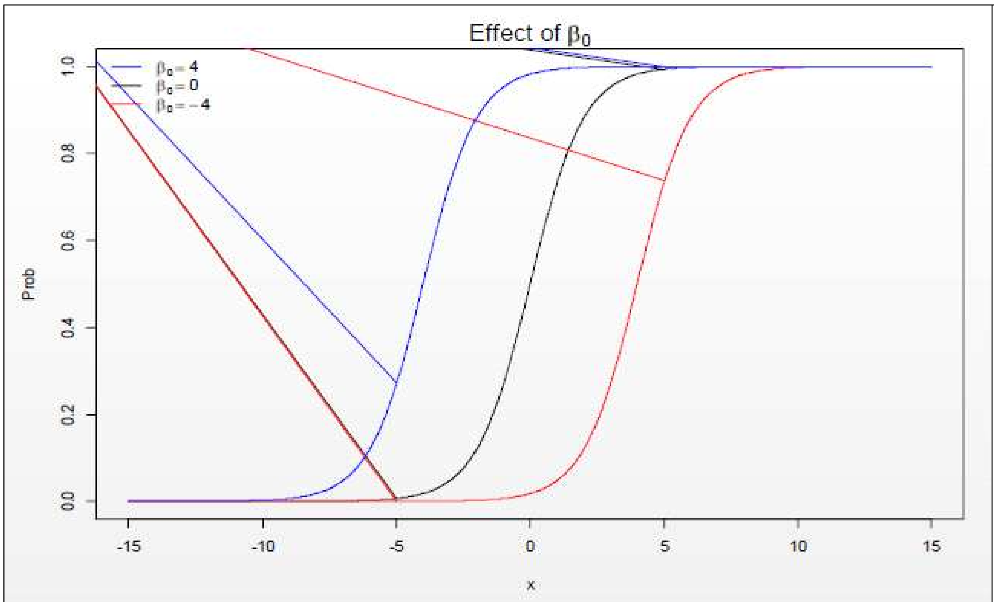
Logit Link Function에 의한 GLM : Logistic Regression : $\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Logistic Regression

이항 반응변수에 logit link function을 적용시킨 GLM
-> $\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

회귀식 : $P(Y=1)$에 관하여 정리
-> $P(Y=1) = \frac{e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$

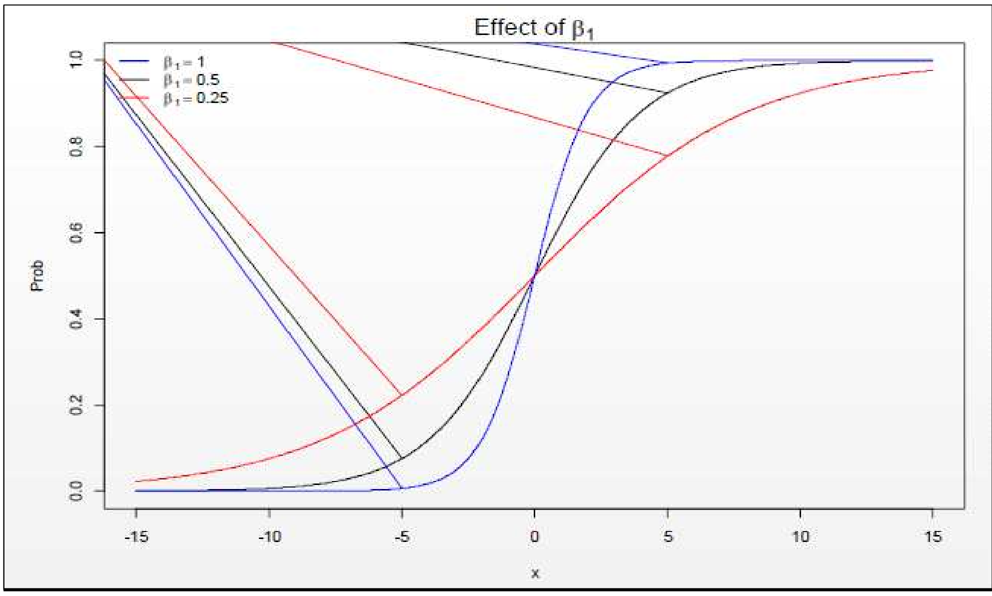
Logistic 회귀식의 특성 : 절편의 효과



$P(Y=1) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}}$

β_0가 증가함에 따라 왼쪽으로 이동
-> 고정된 X 수준에서 확률 증가

Logistic 회귀식의 특성 : 기울기의 효과



$P(Y=1) = \dfrac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}}$

β_1가 증가함에 따라 기울기 증가

Link Function 2 : Probit link

성공확률 : $P(Y=1) = \pi$

Probit Fuction : $\Phi^{-1}(\pi)$, 단, $\Phi(x)$는 누적 표준정규분포 함수

Probit Link Function에 의한 GLM : $\Phi^{-1}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

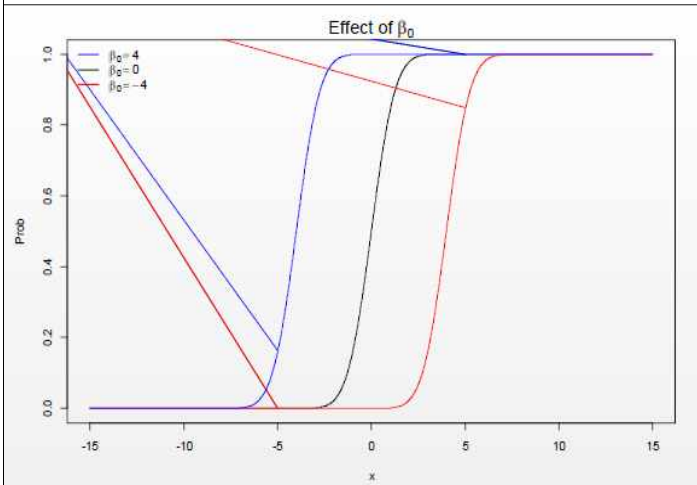
Probit 모형식 : $P(Y=1) = \pi$ 에 관하여 정리

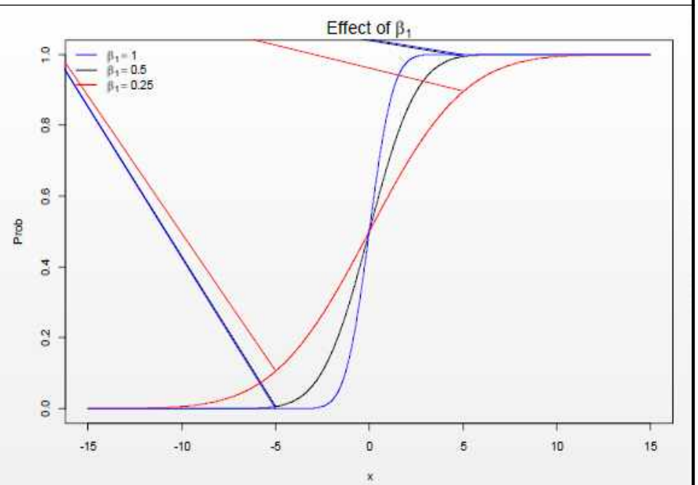
$P(Y=1) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$

Probit 모형식의 특성

$P(Y=1) = \Phi(\beta_0 + \beta_1 X_1)$

Logit 모형식의 특성과 동일





Link Function의 선택 : Logit vs Probit

$\pi \approx 0$ 또는 $\pi \approx 1$ 영역을 제외하면 거의 비슷한 형태

Probit 모형이 더 앞서 도입되었으나 최근에는 Logit 모형이 더 선호된다.

Logit 모형의 장점 : 해석상의 편리함. Odds 활용 가능

Φ에 비해 수학적 처리가 단순하다.

Logistic 회귀모형의 추정 및 해석

로지스틱 회귀 모형 : $\log(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

모수 $\beta_0, \beta_1, \cdots, \beta_p$ 추정 : Maximum Likelihood Estimation (MLE)

- 정규분포의 경우와는 다르게 정확한 $\hat{\beta}$을 구할 수 없음.
- 비선형 정규방정식 → 반복 계산에 의한 추정

모수 추정에 실패하는 경우

설정된 모형이 적절하다면 몇 번의 반복만으로도 모수 추정 가능
반복 계산 수렴 기준을 충족시키지 못해 추정에 실패하는 경우 발생 가능
- 관측값의 크기가 충분히 크지 않았을 때
- 독립변수의 측정 척도가 매우 다를 때
- 성공 혹은 실패 중 한 범주의 발생 빈도가 매우 낮을 때

Logistic 회귀모형 추정을 위한 R 함수

GLM을 위한 R 함수 : glm()
이항 반응변수인 경우 함수 glm()의 일반적인 사용법
- glm(fomular, family=binomail, data, ...)
- fomular : 숫자형 벡터 혹은 요인 (첫 번째 범주가 "실패", 두 번째 범주가 "성공"으로 처리됨.)
- family : 반응변수의 분포 및 link function
    - 이항 반응변수 : binomial
    - link function : 디폴트는 logit (생략됨.)
    - probit을 원하는 경우 : family=binomial(link= "probit")

예 : 부인 직업 참여 여부 결정에 대한 로지스틱 회귀모형 분석

"no" : 첫 번째 범주. "실패" 로 인식
"yes" : 두 번째 범주. "성공" 으로 인식

-> 함수 glm() : "성공" 확률 P(lfp= "yes") 추정

```
> library(carData)
> with(Mroz, table(lfp))
lfp
 no yes
325 428
```

추정 결과 확인

```
> fit1 <- glm(lfp~., family=binomial, Mroz)
> summary(fit1)
Call:
glm(formula = lfp ~ ., family = binomial, data = Mroz)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1062  -1.0900   0.5978   0.9709   2.1893
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27
Number of Fisher Scoring iterations: 4
```

p-값을 정규분포에서 계산

2장에서는 카이제곱 분포에서 계산

$Z \sim N(0,1)$, $Z^2 \sim \chi^2$

변수 wcyes와 hcyes는 가변수

Number of Fisher Scoring iterations: 4 -> 반복 계산 횟수

모형에서 비유의적인 변수(k618,hc) 제거

```
> fit2 <- glm(lfp~.-k618-hc, family=binomial, Mroz)
> fit2 <- update(fit1, .~.-k618-hc)
> summary(fit2)
Call:
glm(formula = lfp ~ k5 + age + wc + lwg + inc, family = binomial,
    data = Mroz)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0428  -1.0853   0.6015   0.9697   2.1801
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.90193    0.54290   5.345 9.03e-08 ***
k5          -1.43180    0.19320  -7.411 1.25e-13 ***
age         -0.05853    0.01142  -5.127 2.94e-07 ***
wcyes        0.87237    0.20639   4.227 2.37e-05 ***
lwg          0.61568    0.15014   4.101 4.12e-05 ***
inc         -0.03367    0.00780  -4.317 1.58e-05 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  906.46  on 747  degrees of freedom
AIC: 918.46
Number of Fisher Scoring iterations: 3
```

추정된 로지스틱 회귀곡선

# 모든 설명변수 포함	# 비유의적 설명변수(k618, hc) 제외
$\hat{\pi}(x) = \dfrac{\exp(3.18 - 1.46k5 - \cdots - 0.03inc)}{1 + \exp(3.18 - 1.46k5 - \cdots - 0.03inc)}$	$\hat{\pi}(x) = \dfrac{\exp(2.9 - 1.43k5 - \cdots - 0.03inc)}{1 + \exp(2.9 - 1.43k5 - \cdots - 0.03inc)}$

직업참여 확률, $P(lfp = yes)$의 추정

\# 함수 predict()에 의한 확률 추정

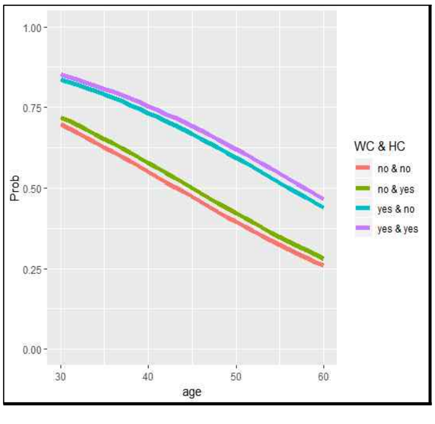
- predict(object, newdata=, type= "response")
- object : 함수 glm()으로 생성된 객체
- newdata = : 새로운 설여변수 값으로 구성된 데이터 프레임. 생략 시 기존 자료에 대한 확률 추정
- type= "response" : 반응변수의 scale로 추정 -> $P(lfp = yes)$의 추정

새로운 설명변수 값에 대한 직업 참여 확률 추정

\# k5, k618, lwg, inc : 평균값　　　age : 30~60　　　wc, hc : 4가지 조합

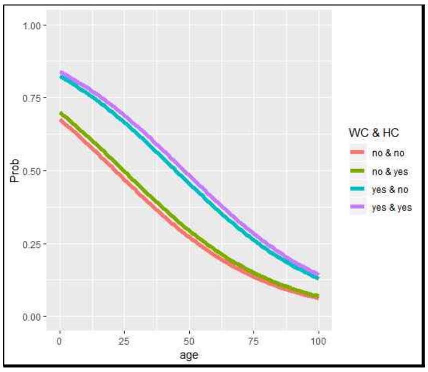
\# new data 만들기

```
> library(tidyverse)
> df1 <- summarize(Mroz,k5=mean(k5),k618=mean(k618),lwg=mean(lwg),inc=mean(inc))
> df1 <- cbind(df1,age=30:60)
# 예측
> prob_1 <- predict(fit1,newdata=cbind(df1,wc="no",hc="no"),type="response")
> prob_2 <- predict(fit1,newdata=cbind(df1,wc="no",hc="yes"),type="response")
> prob_3 <- predict(fit1,newdata=cbind(df1,wc="yes",hc="no"),type="response")
> prob_4 <- predict(fit1,newdata=cbind(df1,wc="yes",hc="yes"),type="response")
> df_2 <- tibble(age=30:60, p1=prob_1, p2=prob_2, p3=prob_3, p4=prob_4)
# 그래프 그리기
> ggplot(data=df_2) +
  geom_line(mapping=aes(x=age,y=p1,col="no & no"),size=2) +
  geom_line(mapping=aes(x=age,y=p2,col="no & yes"),size=2) +
  geom_line(mapping=aes(x=age,y=p3,col="yes & no"),size=2) +
  geom_line(mapping=aes(x=age,y=p4,col="yes & yes"),size=2) + ylim(0,1) + labs(y="Prob",col="WC & HC")
```



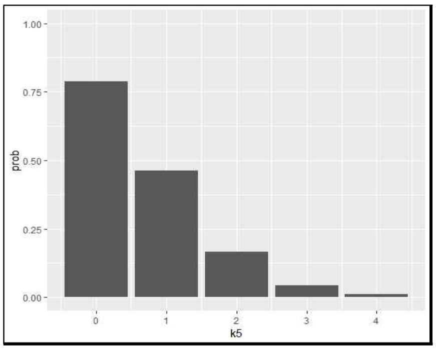
새로운 설명변수 값에 대한 직업 참여 확률 추정

```
# k5, k618, age, lwg : 평균값            inc : 0~100            wc, hc : 4가지 조합

# new data 만들기
〉 df3 〈- summarize(Mroz,k5=mean(k5),k618=mean(k618),age=mean(age),lwg=mean(lwg))
〉 df3 〈- cbind(df3,inc=0:100)
# 예측
〉 prob_1 〈- predict(fit1,newdata=cbind(df3,wc="no",hc="no"),type="response")
〉 prob_2 〈- predict(fit1,newdata=cbind(df3,wc="no",hc="yes"),type="response")
〉 prob_3 〈- predict(fit1,newdata=cbind(df3,wc="yes",hc="no"),type="response")
〉 prob_4 〈- predict(fit1,newdata=cbind(df3,wc="yes",hc="yes"),type="response")
〉 df_4 〈- tibble(age=0:100, p1=prob_1, p2=prob_2, p3=prob_3, p4=prob_4)
# 그래프 그리기
〉 ggplot(data=df_4) +
  geom_line(mapping=aes(x=age,y=p1,col="no & no"),size=2) +
  geom_line(mapping=aes(x=age,y=p2,col="no & yes"),size=2) +
  geom_line(mapping=aes(x=age,y=p3,col="yes & no"),size=2) +
  geom_line(mapping=aes(x=age,y=p4,col="yes & yes"),size=2) + ylim(0,1) + labs(y="Prob",col="WC & HC")
```



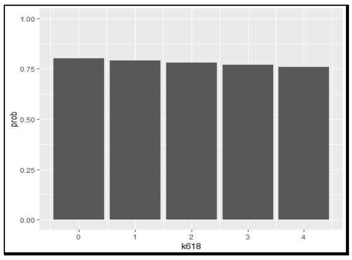
새로운 설명변수 값에 대한 직업 참여 확률 추정

```
# k618, age, lwg, inc : 평균값            k5 : 0~4            wc : “yes”, hc= “yes”

# new data 만들기
〉 df5 〈- summarize(Mroz,k618=mean(k618),age=mean(age),lwg=mean(lwg),inc=mean(inc),wc="yes",hc="yes")
〉 df5 〈- cbind(df5,k5=0:4)
# 예측
〉 prob_1 〈- predict(fit1,newdata=df5,type="response")
〉 df_6 〈- tibble(k5=0:4,p1=prob_1)
# 그래프 그리기
〉 ggplot(data=df_6) +
  geom_bar(mapping=aes(x=k5,y=p1),stat="identity") + ylim(0,1) + labs(y="prob")
```



새로운 설명변수 값에 대한 직업 참여 확률 추정

```
# k5=0            k618=0~4            age,lwg,inc : 평균값            wc : “yes”, hc= “yes”

# new data 만들기
〉 df7 〈- summarize(Mroz,k5=0,age=mean(age),lwg=mean(lwg),inc=mean(inc),wc="yes",hc="yes")
〉 df7 〈- cbind(df7,k618=0:4)
# 예측
〉 prob_1 〈- predict(fit1,newdata=df7,type="response")
〉 df_8 〈- tibble(k618=0:4,p1=prob_1)
# 그래프 그리기
〉 ggplot(data=df_8) +
  geom_bar(mapping=aes(x=k618,y=p1),stat="identity") + ylim(0,1) + labs(y="prob")
```



설명변수의 효과분석

선형회귀모형 : 다른 설명변수들의 수준을 고정시킨 상태에서 X_j를 한 단위 증가시키면 $E(Y)$는 β_j 만큼 변화

로지스틱 회귀모형 : 비선형 모형이기 때문에 선형회귀모형의 방식으로분 효과분석 불가능
대안 1) 확률의 부분변화
대안 2) 확률의 이산변화
대안 3) Odds ratios

Odds ratio에 의한 설명변수 효과 분석

로지스틱 회귀모형 : log(odds)에 대한 모형

-> $\log(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Odds에 대한 모형

-> $\Omega(x) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) = e^{\beta_0} e^{\beta_1 X_1} \cdots e^{\beta_j X_j} \cdots e^{\beta_p X_p}$

설명변수 X_j의 수준을 δ 만큼 변화시켰을 때 odds

-> $\Omega(x, X_j + \delta) = e^{\beta_0} e^{\beta_1 X_1} \cdots e^{\beta_j (X_j + \delta)} \cdots e^{\beta_p X_p}$

설명변수 X_j의 수준을 δ 만큼 변화시켰을 때 odds의 변화 : Odds ratio

$\frac{\Omega(x + X_j + \delta)}{\Omega(x)} = e^{\beta_j \delta}$: 변수 X_j의 효과

예 : Odds ratio에 의한 설명변수의 효과분석

로지스틱 회귀모형 및 회귀계수

```
> fit1 <- glm(lfp~., family=binomial, data=mroz)
> coef(fit1)
(Intercept)          k5         k618         age        wcyes        hcyes          lwg          inc
 3.18214046 -1.46291304 -0.06457068 -0.06287055  0.80727378  0.11173357  0.60469312 -0.03444643
```

log(Odds)에 대한 모형 : 지수 변환으로 Odds에 대한 모형으로 변환

각 설명변수의 Odds ratio 계산

```
> exp(coef(fit1))
(Intercept)         k5       k618        age      wcyes      hcyes        lwg        inc
 24.0982799  0.2315607  0.9374698  0.9390650  2.2417880  1.1182149  1.8306903  0.9661401
```

Odds ratio에 대한 대략적인 해석

- 공통 가정 : 다른 설명변수의 수준은 고정
- 1보다 작은 값 : 해당 설명변수를 1단위 증가 시켰을 때 부인이 직업을 가질 odds 감소
- 1보다 큰 값 : 해당 설명변수를 1단위 증가 시켰을 때 부인이 직업을 가질 odds 증가
- odds 증감은 확률의 증감을 의미

각 설명변수 odds ratio 값에 대한 구체적인 해석

- 다른 설명변수의 수준을 고정시켰을 때
- k5를 한 단위 증가시키면 직업에 참여할 odds ratio는
  $\exp(\hat{\beta}_1) = \exp(-1.4629) = 0.232$배 감소
  $100 \times (0.232 - 1) = -76.8,$ 즉 76.8% 감소
- k5를 두 단위 증가시키면 직업에 참여할 odds ratio는
  $\exp(\hat{\beta}_1 \times 2) = \exp(-1.4629 \times 2) = 0.0536$배 감소
  $100 \times (0.0536 - 1) = -94.6,$ 즉 94.6% 감소

각 설명변수의 Odds ratio 계산

```
> exp(coef(fit1))
(Intercept)         k5       k618        age      wcyes      hcyes        lwg        inc
 24.0982799  0.2315607  0.9374698  0.9390650  2.2417880  1.1182149  1.8306903  0.9661401
```

- 다른 설명변수의 수준을 고정시켰을 때
- lwg를 한 단위 증가시키면 직업에 참여할 odds ratio는
  $\exp(\hat{\beta}_6) = \exp(0.6047) = 1.831$배 증가
  $100 \times (1.831 - 1) = 83.1,$ 즉 83.1% 증가
- lwg를 두 단위 증가시키면 직업에 참여할 odds ratio는
  $\exp(\hat{\beta}_6 \times 2) = \exp(0.6047 \times 2) = 3.35$배 증가
  $100 \times (3.35 - 1) = 235.1,$ 즉 235.1% 증가
- 부인 학력수준(wc)이 대졸인 경우와 고졸 이하의 경우와 비교하여 직업에 참여할 odds ratio는 2.242배 증가
  $100 \times (2.242 - 1) = 124.2,$ 즉 124.2% 증가

각 설명변수의 Odds ratio에 대한 95% 신뢰구간

```
> exp(confint(fit1))
Waiting for profiling to be done...
                 2.5 %      97.5 %
(Intercept) 6.9377228 87.0347916
k5          0.1555331  0.3370675
k618        0.8200446  1.0710837
age         0.9154832  0.9625829
wcyes       1.4347543  3.5387571
hcyes       0.7467654  1.6766380
lwg         1.3689201  2.4768235
inc         0.9502809  0.9814042
```

\# 신뢰구간에 1이 포함되어 있는 변수
- 비유의적 변수
- summary(fit1) 결과와 비교

\# profile likelihood 방식에 의한 신뢰구간 계산 :
- Wald 검정 방식에 의한 교재 표2.12의 결과와는 약간 다름
- 신뢰구간이 odds ratio 점추정값에 대하여 좌우대칭이 아님.

"가설검정" 에서 신뢰구간에 대한 추가 설명 예정

Probit 모형

\# Probit 모형 : $\Phi^{-1}[P(Y=1)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

함수 $\Phi(x)$는 누적 표준정규분포

\# 추정된 probit 모형 : $P(Y=1) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$

예 : 직업 참여자료에 대한 Probit 모형 적합

```
> fit.p <- glm(lfp~., family=binomial(link="probit"), data=Mroz)
> summary(fit.p)
Call:
glm(formula = lfp ~ ., family = binomial(link = "probit"), data = Mroz)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1359  -1.1024   0.5967   0.9746   2.2236
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.918418   0.382356   5.017 5.24e-07 ***
k5          -0.874712   0.114423  -7.645 2.10e-14 ***
k618        -0.038595   0.040950  -0.942 0.345942
age         -0.037824   0.007605  -4.973 6.58e-07 ***
wcyes        0.488310   0.136731   3.571 0.000355 ***
hcyes        0.057172   0.124207   0.460 0.645306
lwg          0.365635   0.089992   4.063 4.85e-05 ***
inc         -0.020525   0.004852  -4.230 2.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.39  on 745  degrees of freedom
AIC: 921.39
Number of Fisher Scoring iterations: 4
```

회귀모형의 회귀계수 추정값 비교 (logit vs probit)

```
> cbind(logit=round(coef(fit1),3),probit=(round(coef(fit.p),3)))
             logit probit
(Intercept)  3.182  1.918
k5          -1.463 -0.875
k618        -0.065 -0.039
age         -0.063 -0.038
wcyes        0.807  0.488
hcyes        0.112  0.057
lwg          0.605  0.366
inc         -0.034 -0.021
```

기존 자료에 대한 직업 참여 확률 추정 비교 (logit vs probit)

```
> cbind(logit=fit1$fitted.values,probit=fit.p$fitted.values)[1:10,]
      logit     probit
1  0.5158291 0.5206967
2  0.6668165 0.6650898
3  0.4565831 0.4643790
4  0.6620169 0.6593693
5  0.6632299 0.6653360
6  0.5959744 0.5958797
7  0.9242061 0.9354251
8  0.6586118 0.6573715
9  0.4738387 0.4785964
10 0.7483850 0.7471961
```

거의 동일한 결과
회귀계수의 차이는 모형의 다름으로 인한 것

probit 모형의 단점 : 개별 설명변수의 효과분석에서 로지스틱 회귀모형과는 다르게 odds ratio에 의한 분석 불가능 -> 상당한 불편함을 초래

적용분야

[적용 분야] 로지스틱 회귀분석의 주요 목적 : 판별분석과 거의 동일
- 반응변수의 구분을 설명할 수 있는 모형 추정 : 두 가지 명목형 범주의 차이를 설명할 수 있는 비선형 모형 추정
- 각 범주에 속할 확률 추정 : 추정된 모형을 근거로 주어진 설명변수 수준에서 각 범주에 속할 확률 추정
- 범주에 대한 분류 : 추정된 확률을 근거로 각 관찰값의 범주를 예측

[적용 예]
- 중소기업의 부실 여부 예측
- 신상품 구매의사 성향 예측
- 특정 질환 판정 예측
- 보험 부당 청구 탐지