3. 일반화 선형모형(Generalized Linear Model)

내용

1. 일반화 선형모형(GLM)에 대한 소개
2. 이항 반응변수에 대한 선형회귀모형

1. 일반화 선형모형(GLM)의 소개

- 통상적인 선형회귀모형
  - 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 선형회귀모형(Classical Linear Regression model)의 한계점

- 모형: $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i$

- 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

GLM의 세가지 성분

- Random component
- Systematic component
- Link function

$$Y \xleftrightarrow{\text{Link Function}} \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

Random Component

Systematic Component

1) Random Component

- 반응변수 Y의 확률분포 규정
  - GLM에서 반응변수 Y의 분포는 Exponential family에 속해야 한다.
  - Exponential family에 속하는 분포 예:
    정규분포, 포아송분포, 이항분포, 감마분포 등등

2) Systematic Component

- 반응변수에 대한 설명변수의 영향력을 표현

설명변수($x_1, \cdots, x_p$)의 선형결합(Linear predictor)

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

3) 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, $\log(\mu/(1-\mu)) = \eta$

2. 이항 반응변수에 대한 선형회귀모형

- 이항 반응변수: 두 가지 범주만을 갖는 변수. 일반적으로 1 혹은 0의 값을 부여한다.
- 이항 반응변수의 분포: Bernoulli 분포

$$\mathrm{P}(Y = y) = \pi^{y}(1-\pi)^{1-y}, \qquad y = 0, 1$$

- 이항 반응변수의 평균과 분산

$$\mathrm{E}(Y) = \sum_{y} y \times P(Y = y) = P(Y = 1) = \pi$$

$$\mathrm{Var}(Y) = E(Y^2) - \big(E(Y)\big)^2 = \pi(1-\pi)$$

이항 반응변수에 대한 선형회귀모형설정 1

- Classical Linear Regression Model: GLM with identity link

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

  - $Y_i = 0, 1$ → 오차항의 가정을 만족시킬 수 없음
  - Random component와 systematic component의 범위가 다름

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

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

    이항반응변수에 대해서는 Classical linear regression model을
    적용시킬 수 없음

예제 2.1

- `carData::Mroz` : 결혼한 미국 백인 여성의 직업참여 여부 분석
- 반응 변수: `lfp`(labor-force participation): no, yes
- 설명변수

  `k5`: 5세 이하 자녀 수

  `k618`: 6~18세 자녀 수

  `age`: 부인 나이

  `wc`: 부인 대학 교육 여부 (no, yes)

  `hc`: 남편 대학 교육 여부 (no, yes)

  `lwg`: 부인 기대 소득의 로그 값

  - 직업이 없는 경우, 다른 변수를 이용한 예측 값

  `inc`: 부인 소득을 제외 가계 소득

(1) 자료 Mroz의 구조

```
> 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 ...
```

- 자료 Mroz에 있는 변수들의 요약 통계량 계산

```
> summary(Mroz)
  lfp          k5              k618             age            wc
 no :325   Min.   :0.0000   Min.   :0.000   Min.   :30.00   no :541
 yes:428   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   yes:212
           Median :0.0000   Median :1.000   Median :43.00
           Mean   :0.2377   Mean   :1.353   Mean   :42.54
           3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00
           Max.   :3.0000   Max.   :8.000   Max.   :60.00
   hc         lwg               inc
 no :458   Min.   :-2.0541   Min.   :-0.029
 yes:295   1st Qu.: 0.8181   1st Qu.:13.025
           Median : 1.0684   Median :17.700
           Mean   : 1.0971   Mean   :20.129
           3rd Qu.: 1.3997   3rd Qu.:24.466
           Max.   : 3.2189   Max.   :96.000
```

(2) 선형회귀모형 추정 및 검정

- 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

```
> library(dplyr)
> 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
```

- 선형회귀모형 적합

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

- 매우 낮은 결정계수
- 회귀계수는 유의함

```
> fit1 <- lm(lfp~k5, data=mroz)
> summary(fit1)

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 ***

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
```

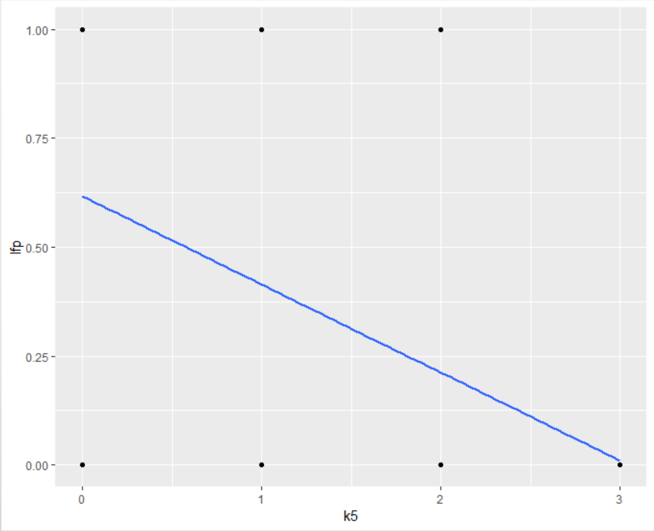
- 추정된 회귀직선 및 반응변수의 관찰값

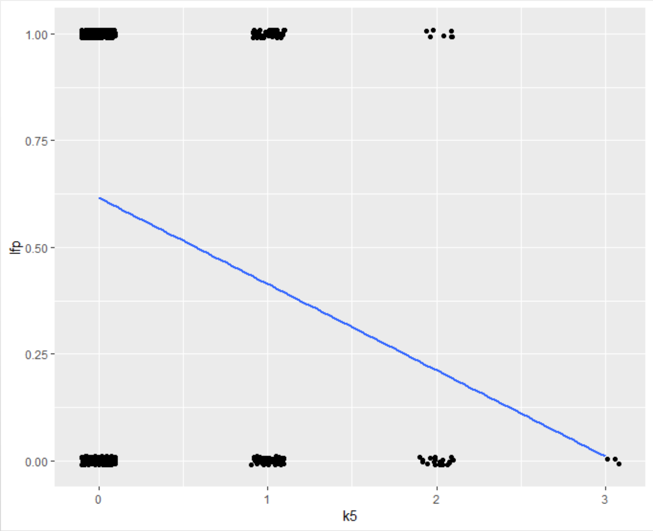
```
> library(ggplot2)
> ggplot(mroz, aes(x=k5, y=lfp)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE)
```

총 관찰값의 개수는 753개
- 그래프에는 7개의 점만이 나타남
- 중복으로 인한 결과

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

- jitter: 점의 위치에 random noise 추가





- 추정된 회귀모형의 문제점

  - 5세 이하 자녀의 수(k5)가 증가함에 따라 부인이 직업을 가질 확률은 감소
  - k5=4인 경우, 확률값이 **음수**로 추정

    → 회귀모형의 적합성에 중대한 문제

k5=4인 경우의 적합값 예측(95% 예측 구간 포함)

```
> predict(fit1, newdata=data.frame(k5=4), interval="confidence")
         fit        lwr        upr
1 -0.1923117 -0.4437612 0.05913788
```

(3) 모든 설명변수 포함된 회귀모형 적합

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

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 ***

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

→ 잘못 설정된 회귀모형의 함수 형태가 원인일 가능성이 높음

이항 반응변수에 대한 선형회귀모형설정 2

- 일반화 선형모형(GLM) 적용

  - Random component: 반응변수 Y의 분포
    Bernoulli 분포는 Exponential family에 속함

  - 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 g선택. 단, $0 \le g^{-1}(\eta_i) = \pi_i \le 1$를 만족

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

- Link Function 1 : Logit link

  - 성공 확률: $P(Y = 1) = \pi$
  - Odds: $\Omega = P(Y = 1)/1 - P(Y = 1), \ \ 0 \le \Omega \le \infty$
  - Logit Function:

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

  - 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

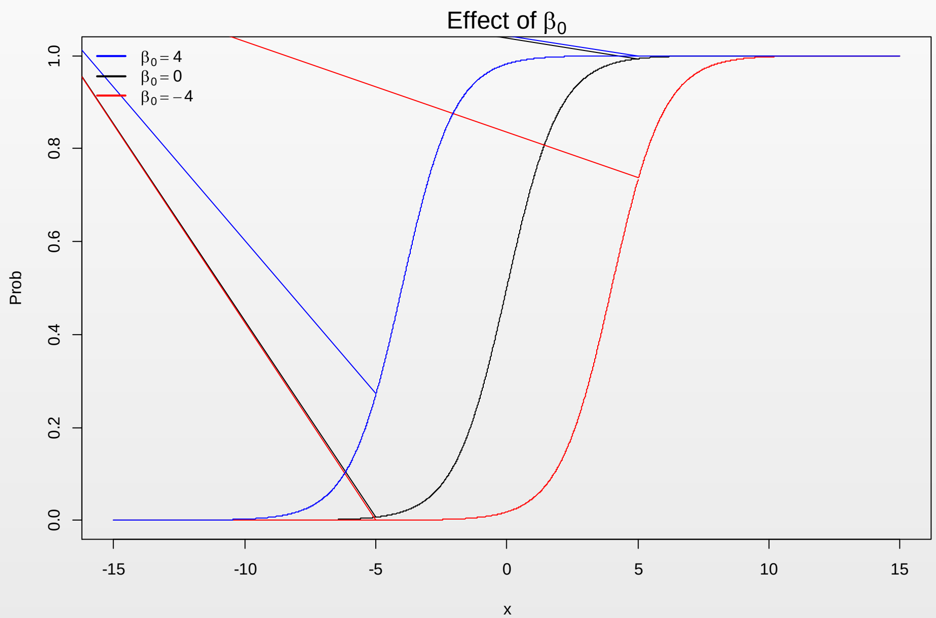
  - 이항 반응변수에 logit link function을 적용시킨 GLM

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

  - 회귀식: $P(Y=1)$에 관하여 정리

$$P(Y=1) = \frac{exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}$$

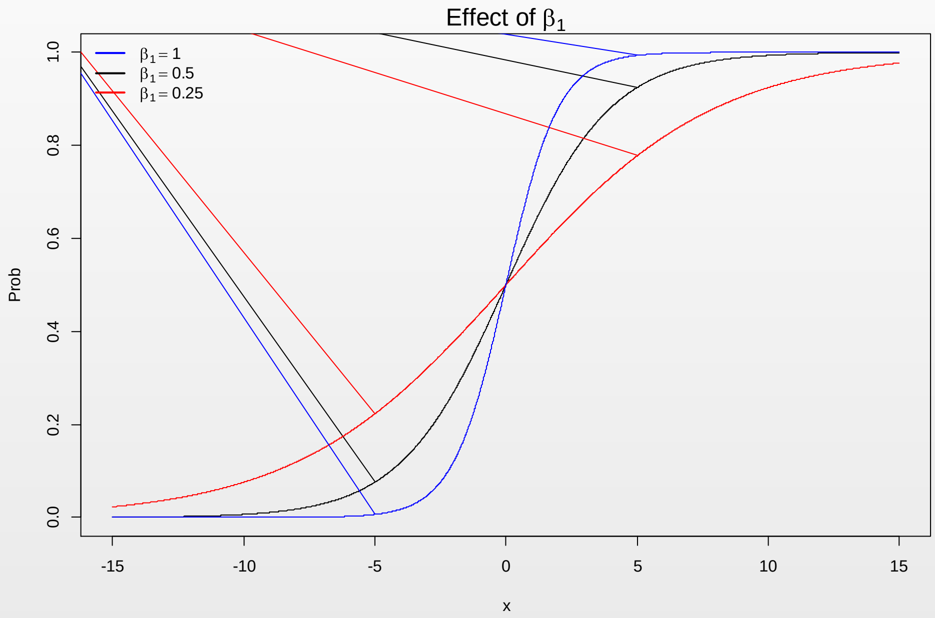
- <u>Logistic 회귀식의 특성: 절편의 효과</u>



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

β_0가 증가함에 따라 왼쪽으로 이동 → 고정된 X 수준에서 확률 증가

- <u>Logistic 회귀식의 특성: 기울기의 효과</u>



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

β_1이 증가함에 따라 기울기 증가

- Link Function 2 : Probit Link

  - 성공 확률: $P(Y=1)=\pi$
  - Probit Function: $\Phi^{-1}(\pi)$, 단 $\Phi(x)$는 누적 표준정규분포 함수
  - Probit Link Function에 의한 GLM:

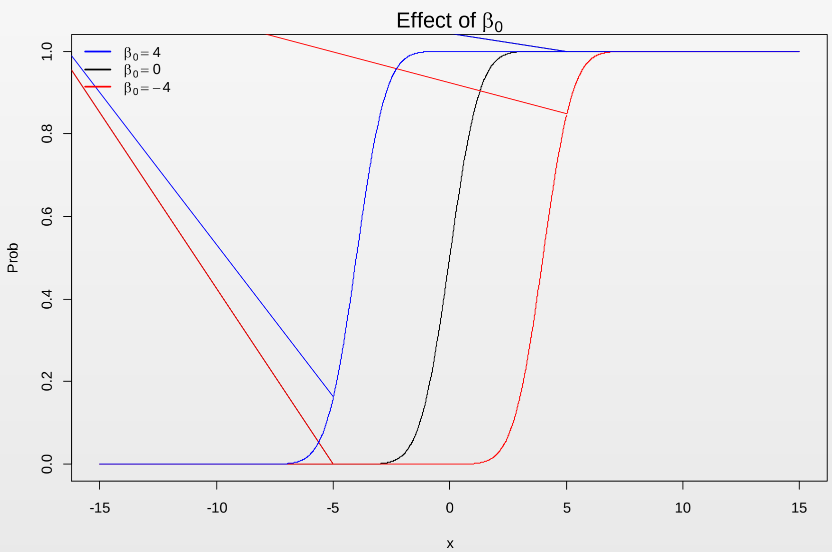
$$\Phi^{-1}(\pi)=\beta_0+\beta_1 X_1+\cdots+\beta_p X_p$$

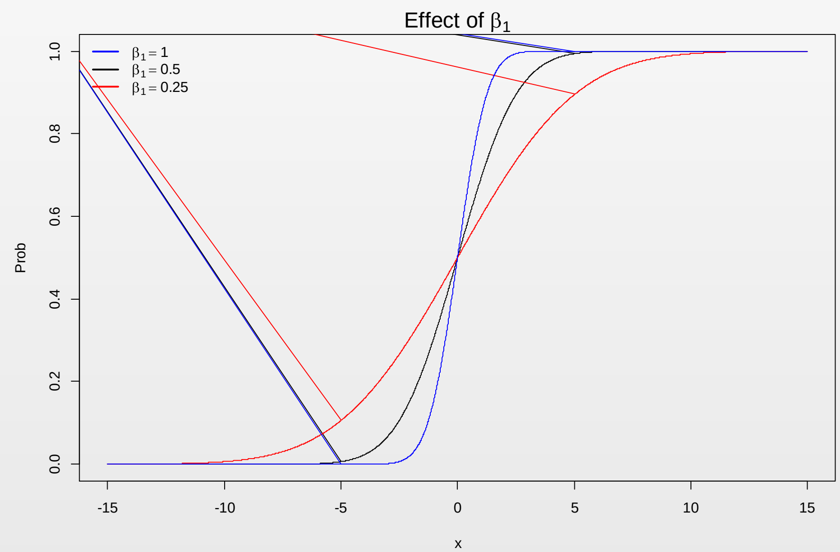
  - Probit 모형식: $P(Y=1)=\pi$에 관하여 정리

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

- Probit 모형식의 특성

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





Logit 모형식의 특성과 동일

- Link Function의 선택: Logit or Probit

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

  - Probit 모형이 더 앞서 도입되었으나 최근에는 Logit 모형이 더 선호됨

  - Logit 모형의 장점
    1) 해석상의 편리함: odds 활용 가능
    2) Φ에 비해 수학적 처리가 단순