12강

일반화선형모형 (1)

정보통계학과 김성수교수

학습목차

1. 일반화선형모형
2. 로지스틱 회귀모형

1 일반화선형모형

선형회귀모형

✓ 회귀모형

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

✓ 회귀모형에 부여되는 가정

- 오차의 등분산성 가정. 모형의 선형성 가정. 오차의 정규성 가정.

✓ 반응변수 Y 가 정규분포가 아닌 경우

- 오차의 등분산성이 위배되는 경우 : 분산안정화변환 등으로 해결
- 오차의 정규성이 위배되는 경우 : 일반화선형모형

일반화선형모형

✓ 반응변수 Y 가 정규분포를 따르지 않는 경우

예) 반응변수가 비율을 나타내는 경우

반응변수가 양의 개수를 나타내는 포아송분포를 따르는 경우

✓ 일반화선형모형(generalized linear model)

- 반응변수의 분포가 정규분포인 경우 뿐만 아니라, 이항분포, 포아송분포, 감마분포, 음이항분포, 역정규분포 등과 같은 지수족(exponential family) 분포를 따른다고 할 때, 회귀모형의 틀에서 통합적으로 확장된 모형
- Nelder와 Wedderburn(1972)이 제안

일반화선형모형

✓ 일반화선형모형의 세가지 구성성분

① 반응변수의 분포 ② 선형예측자 ③ 연결함수

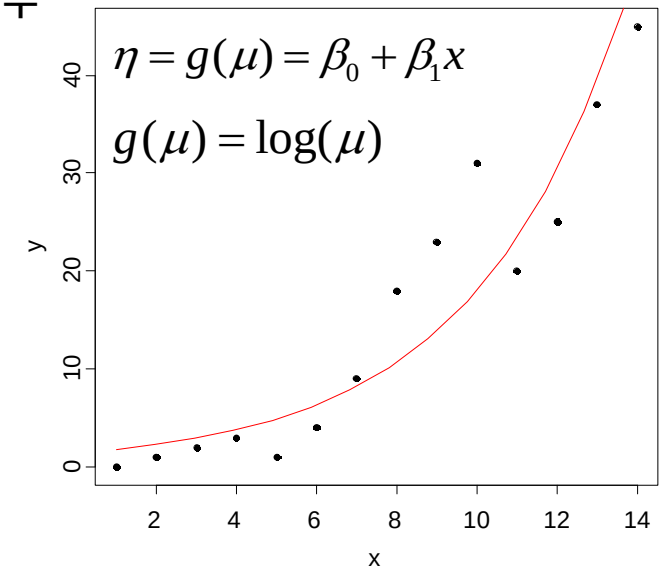
예) AIDS data : Whyte, et.al. 1987 (Dobson, 1990).

1983~1986년 동안 Australia에서 AIDS로 인한 사망자 수

X : 1983년1월 부터 시작한, 3개월 단위 경과기간

Y : 사망자 수

X	Y	X	Y
1	0	8	18
2	1	9	23
3	2	10	31
4	3	11	20
5	1	12	25
6	4	13	37
7	9	14	45



일반화선형모형

✓ 선형모형의 일반화선형모형으로의 확장

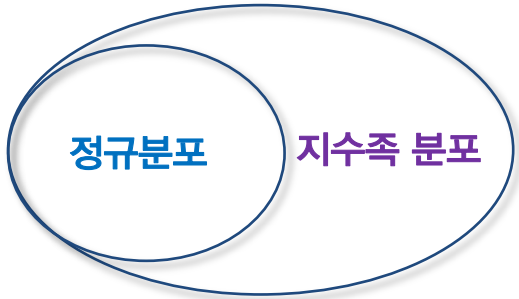
	선형회귀모형	일반화선형모형
반응변수의 분포	정규분포를 가정	정규분포, 이항분포, 포아송 분포, 음이항분포, 감마분포 등 지수족 분포 중 하나를 가정
평균의 선형성	$\mu = E(Y) = X^T\beta$	$\eta = g(\mu) = X^T\beta$
모수 추정법	최소제곱추정 (= 최대가능도추정)	최대가능도추정

반응변수의 분포 : 지수족 분포

✓ **지수족 분포** (the exponential family of distributions) :
확률밀도 함수 $f(y;\theta,\phi)$ 가 다음과 같이 표현되는 분포.

지수족 분포의 확률밀도함수 :

$$f(y;\theta,\phi) = q(y,\phi/w)\exp\left(\frac{y\theta - b(\theta)}{\phi/w}\right)$$



θ : 평균 μ의 함수로 표현되는 정준모수(canonical parameter)

ϕ : y의 분산과 관련된 평균과는 독립인 산포모수
(dispersion parameter)

w : y분포가정에 따라 사전에 알 수 있는 값

반응변수의 분포 : 지수족 분포

✓ **정규분포의 예**

$$Y \sim N(\mu, \sigma^2),\ -\infty < y < \infty,\ -\infty < \mu < \infty,\ 0 < \sigma < \infty.$$

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2}\right)$$

, $\theta = \mu,\ b(\theta) = \frac{1}{2}\theta^2,\ \phi = \sigma^2,\ w = 1$

, $E(Y) = b'(\theta) = \theta = \mu,$

$Var(Y) = \frac{\phi}{w} b''(\theta) = \sigma^2$

반응변수의 분포 : 지수족 분포

✓ **이항분포의 예**

$$Y \sim B(n,\pi),\ y=0,1,\cdots,n,\ 0<\pi<1,$$

$$f(y;\pi)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y}=\binom{n}{y}exp[y\theta-n\log(1+e^{\theta})]$$

: $\theta=\log\left(\frac{\pi}{1-\pi}\right),\ b(\theta)=n\log(1+e^{\theta}),\ \phi=1,\ w=1$

: $E(Y)=b'(\theta)=\frac{ne^{\theta}}{1+e^{\theta}}=n\pi,$

$$Var(Y)=\frac{\phi}{w}b''(\theta)=\frac{ne^{\theta}}{(1+e^{\theta})^{2}}=n\pi(1-\pi)$$

반응변수의 분포 : 지수족 분포

✓ **포아송분포의 예**

$$Y \sim poisson(\mu),\ y = 0,1,2,\cdots,\ 0 < \mu < \infty$$

$$f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!} = \frac{1}{y!} exp(y\theta - e^{\theta})$$

, $\theta = \log\mu,\ b(\theta) = e^{\theta},\ \phi = 1,\ w = 1$

, $E(Y) = b'(\theta) = e^{\theta} = \mu,$

$$Var(Y) = \frac{\phi}{w} b''(\theta) = e^{\theta} = \mu$$

선형예측자와 연결함수

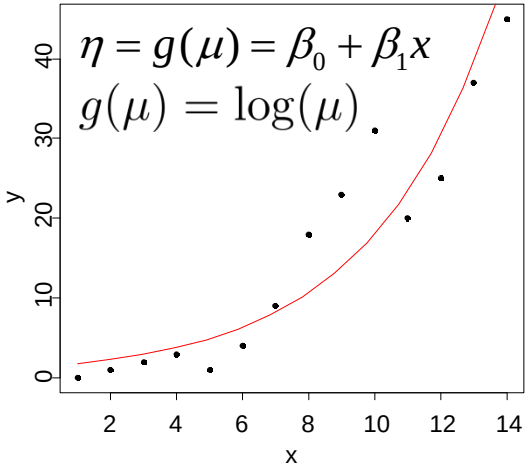
✓ 선형예측자(linear predictor)

모형에 있는 설명변수들의 선형결합을 선형예측자라고 함.

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

✓ 연결함수(link function)

선형예측자 $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 와 반응변수의 평균 μ 사이의 관계를 $g(\mu) = \eta$ 가 되도록 연결해주는 함수 $g(\bullet)$



선형예측자와 연결함수

✓ 지수족 분포의 정준연결

반응변수의 분포	평균의 모수 공간	정준연결	이름	평균함수
정규분포	$(-\infty, \infty)$	$g(\mu) = \mu$	항등함수	$\mu = x'\beta$
베르누이분포	$(0,1)$	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	로짓함수	$\mu = \frac{\exp(x'\beta)}{\exp(1+x'\beta)}$
포아송분포	$(0,\infty)$	$g(\mu) = \log(\mu)$	로그함수	$\mu = \exp(x'\beta)$
감마분포	$(0,\infty)$	$g(\mu) = \frac{1}{\mu}$	역수함수	$\mu = (x'\beta)^{-1}$

2 로지스틱 회귀모형

로지스틱 회귀모형 : 이항자료

〈날다람쥐 Sugar Glider의 출현자료〉

p_no	occurr	con_metric	p_size_km
1	1	0.650	130.9
2	0	0.610	104.1
3	0	0.744	132.3
4	1	0.213	225.6
5	1	0.723	83.0
6	0	0.678	48.8
7	0	0.733	61.0
8	1	0.522	39.6
9	1	0.552	193.1
10	0	0.245	155.6

```
> glider <-
    read.csv('c:/data/reg/sugar_glider_binomial.csv')
> head(glider, 3)
  p_no occurr con_metric p_size_km
1    1      1      0.650     130.9
2    2      0      0.610     104.1
3    3      0      0.744     132.3
> logit_m1 <- glm(occurr~p_size_km+con_metric,
                  family=binomial(link=logit),
                   data=glider)
```

반응변수 y=occur, 1=yes, 0=no **이므로 이항분포를 가정**

로지스틱 회귀모형 :

$$\eta = \log it(\pi) = \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\pi = E(Y \mid x) = \Pr(y = 1 \mid x)$$

R 결과

```
> logit_m1 <- glm(occurr~p_size_km+con_metric, family=binomial(link=logit), data=glider)
> summary(logit_m1)
...
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4969  -0.8829  -0.3884   0.8766   2.0515

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.606207   1.436391  -2.511  0.01205 *
p_size_km    0.023566   0.007462   3.158  0.00159 **
con_metric   1.631800   1.642758   0.993  0.32055
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 54.661  on 47  degrees of freedom
AIC: 60.661
Number of Fisher Scoring iterations: 4
```

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2=-3.606+0.024^* x_1+1.632^* x_2$$

Deviance(이탈도) : 선형회귀모형의 잔차 제곱합을 일반화한 개념. 정규분포를 따르는 가정이 맞는 경우, χ^2- 분포를 따름.

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_0$$

모형의 이탈도

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2$$

R 결과 : 모형의 유의성 검정

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

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 54.661  on 47  degrees of freedom
AIC: 60.661
Number of Fisher Scoring iterations: 4
```

모형의 유의성 검정

$$H_0 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 \quad vs. \quad H_1 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

즉,

$$H_0 : \beta_1 = 0,\ \beta_2 = 0 \quad vs. \quad H_1 : \text{적어도 하나는 } 0\text{이 아니다.}$$

p-값 계산

```
> 1-pchisq(68.994-54.661,2)
[1] 0.0007720201
```

=> p-값이 매우 작으므로 대립가설의 모형이 유의함을 알 수 있음

R 결과 : anova를 이용한 유의성 검정

```
> 1-pchisq(68.994-54.661,2)
[1] 0.0007720201
> logit_m0 <- glm(occurr~1, family=binomial(link=logit), data=glider)
> anova(logit_m0, logit_m1, test='Chisq')
Analysis of Deviance Table

Model 1: occurr ~ 1
Model 2: occurr ~ p_size_km + con_metric
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        49     68.994
2        47     54.661  2   14.333 0.000772 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

R 결과 : 모형의 적합성 검정

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

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 54.661  on 47  degrees of freedom
AIC: 60.661
Number of Fisher Scoring iterations: 4
```

모형의 적합성 검정

$$H_0 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

p-값 계산

```
> 1-pchisq(54.661, 47)
[1] 0.2064349
```

=> p-값 > 유의수준 0.05 이므로, 모형이 적합하다고 판단. 보통의 경우, "Residual deviance/df < 2" 이면 모형의 적합도에 큰 문제가 없다고 판단함.

모형의 선택

변수 선택 : x_1이 선택 된 모형에서 x_2를 추가하는 것이 적절한 지 검정

$$H_0 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 \quad vs. \quad H_1 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

```
> logit_m2 <- glm(occurr ~ p_size_km, family=binomial(link=logit), data=glider)
> logit_m1 <- glm(occurr~p_size_km+con_metric, family=binomial(link=logit),
+                     data=glider)
> anova(logit_m2, logit_m1, test='Chisq')
Analysis of Deviance Table

Model 1: occurr ~ p_size_km
Model 2: occurr ~ p_size_km + con_metric
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        48     55.716
2        47     54.661  1   1.0546   0.3045
```

=> 검정결과 변수 x_2 를 추가하는 것이 유의하지 않음(p-값 = 0.3045)

모형의 선택 : AIC 함수

변수 선택 : x_1이 선택 된 모형에서 x_2를 추가하는 것이 적절한 지 검정

$$H_0 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 \quad vs. \quad H_1 : \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

```
> AIC(logit_m2, logit_m1)
          df      AIC
logit_m2   2 59.71577
logit_m1   3 60.66120
```

=> 아카이케 정보기준에 근거한 모형평가 : 작은 값을 가지는 모형을 선택

모형의 선택 : 변수선택방법 이용

```
> library(MASS)
> stepAIC(logit_m1, direction='both')
Start:  AIC=60.66
occurr ~ p_size_km + con_metric

             Df Deviance    AIC
- con_metric  1   55.716 59.716
<none>            54.661 60.661
- p_size_km   1   68.889 72.889

Step:  AIC=59.72
occurr ~ p_size_km

             Df Deviance    AIC
<none>            55.716 59.716
+ con_metric  1   54.661 60.661
- p_size_km   1   68.994 70.994
```

```
Call:  glm(formula = occurr ~ p_size_km, family =
binomial(link = logit),
    data = glider)

Coefficients:
(Intercept)    p_size_km
   -2.52830      0.02173

Degrees of Freedom: 49 Total (i.e. Null);  48 Residual
Null Deviance:      68.99
Residual Deviance: 55.72        AIC: 59.72
```

변수 x_1 (p_size_km) 이 선택됨.

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 = -2.528 + 0.022^* x_1$$

로지스틱 모형의 해석

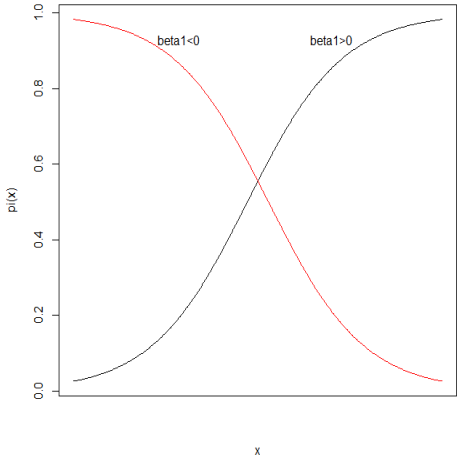
로짓함수

$$\pi(x) = E(Y \mid x) = \Pr(y = 1 \mid x)$$

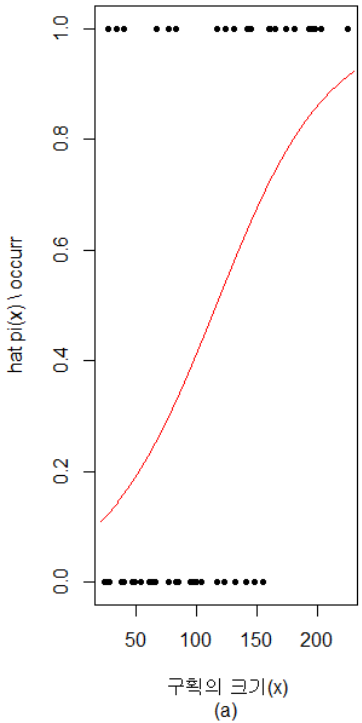
$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

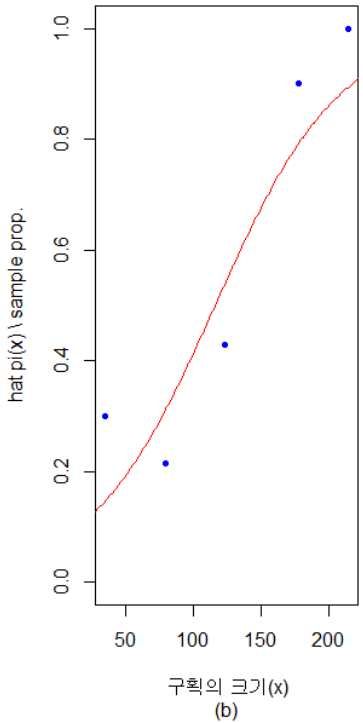
: $\pi(x)$ **는** x **가 증가함에 따라** S **자형 곡선을 그리며,**

$\beta_1 > 0$ **이면 단조증가,** $\beta_1 < 0$ **이면 단조감소함.**



로지스틱 모형의 해석 : $\hat{\pi}(x)$



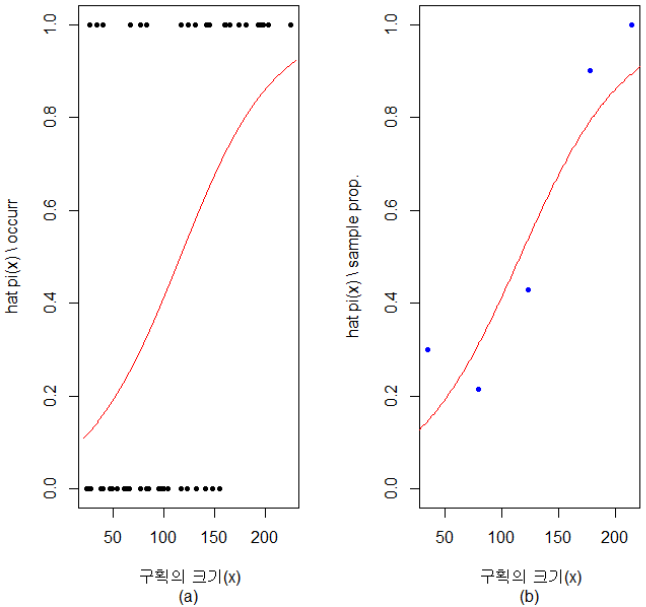


<구획 크기의 계급구간에서 구획 수, Sugar Glider 출현 구획 수, 표본비율>

p_size_km	구간의 중앙값	출현 구획 수	구획 수	표본비율
≦ 50.0	35.3	3	10	0.30
50.0 ~ 100.0	79.55	3	14	0.21
100.0 ~ 150.0	123.6	6	14	0.43
150.0 ~ 200.0	177.65	9	10	0.90
200.0 <	214.55	2	2	1.00

로지스틱 회귀모형으로 추정되는 $\hat{\pi}(x)$는 이들 해당 구간에서의 성공률 $\pi(x)$를 추정한 값으로 해석. :

로지스틱 모형의 해석 : R 코드




```
> p_size <- seq(20, 230, 1)
> hat_eta <- predict(logit_m2, list(p_size_km=p_size),
                     type="link")
> par(mfrow=c(1,2))
> plot(p_size_km, occurr, xlab='구획의 크기(x)',
       ylab="hat pi(x) ₩₩ occurr", sub='(a)', pch=20)
> lines(p_size, exp(hat_eta)/(1+exp(hat_eta)),lwd=1.5,
       col='red')
> glider_g <- read.csv('c:/data/regsugar_glider_binomial_g.csv')
> plot(glider_g$p_size_med, glider_g$cases/glider_g$count,
    xlab='구획의 크기(x)',
    ylim=c(0,1), ylab="hat pi(x) ₩₩ sample prop.", sub='(b)',
    pch=20, col='blue')
> lines(p_size, exp(hat_eta)/(1+exp(hat_eta)), lwd=1.5,
       col='red')
```

정리된 자료의 로지스틱 회귀모형 적합

<구획 크기의 계급구간에서 구획 수, Sugar Glider 출현 구획 수, 표본비율>

p_size_km	구간의 중앙값	출현 구획 수	구획 수	표본비율
≦ 50.0	35.3	3	10	0.30
50.0 ~ 100.0	79.55	3	14	0.21
100.0 ~ 150.0	123.6	6	14	0.43
150.0 ~ 200.0	177.65	9	10	0.90
200.0 <	214.55	2	2	1.00

```
> glider_g <- read.csv('c:/data/reg/sugar_glider_binomial_g.csv')
> head(glider_g)
  p_size_med count cases
1      35.30    10     3
2      79.55    14     3
3     123.60    14     6
4     177.65    10     9
5     214.55     2     2
> y <- cbind(glider_g$cases, glider_g$count-glider_g$cases)
> logit_mg <- glm(y~glider_g$p_size_med, family=binomial(link=logit))
```

정리된 자료의 로지스틱 회귀모형 적합

```
> summary(logit_mg)

Deviance Residuals:
      1       2       3       4       5
 1.2452  -0.7897  -0.8196   0.9238   0.6694

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.539381   0.839355  -3.025  0.00248 **
glider_g$p_size_med   0.021776   0.007073   3.079  0.00208 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16.6058  on 4  degrees of freedom
Residual deviance:  4.1477  on 3  degrees of freedom
AIC: 18.547

Number of Fisher Scoring iterations: 4
```

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)=\hat{\beta}_0+\hat{\beta}_1 x_1 = -2.539+0.022^* x_1$$

원 자료를 이용한 모형

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)=\hat{\beta}_0+\hat{\beta}_1 x_1 = -2.528+0.022^* x_1$$

다음시간 안내

13강. 일반화선형모형 (2)