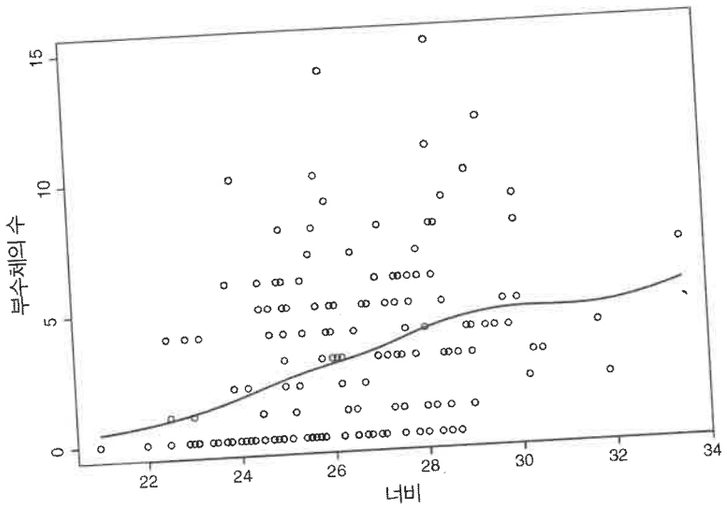

그림 3.3 암컷 참게의 등딱지 너비(cm)에 대한 부수체의 개수 산포도와 일반화가법모형적합에 의한 평활화 곡선

3.3.3 예제: 암참게와 부수체에 관한 연구

〈표 3.2〉에 있는 자료는 참게집에 대한 자료이다[2]. 각 암참게들은 집을 갖고 있으며 이 집에 붙어사는 숫참게를 가지고 있다. 이 숫참게를 **부수체**(附隨體, satellite)라고 부른다[3]. 이 연구에서는 암참게가 부수체를 갖는지 여부에 영향을 미치는 요인을 조사하고 있다. 각 암참게에 대한 반응변수는 부수체의 개수이다. 영향을 줄 가능성이 있을 것으로 생각되는 설명변수의 한 예는 암참게의 크기를 결정하는 등딱지 너비이다.

▸표 3.2 암참게의 등뼈의 상태, 등딱지 너비, 무게에 의한 부수체의 수

C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8

주의: C=색깔(1=약간 밝은색, 2=중간색, 3=약간 어두운색, 4=어두운색), S=등뼈의 상태(1=둘 다 양호함, 2=1개가 휘거나 좋지 않음, 3=둘 다 휘거나 좋지 않음), W=등딱지 너비(cm), Wt=무게(kg), Sa=부수체의 수.

출처: Data provided by Dr. Jane Brockmann, Zoology Department, University of Florida; study described in *Ethology*, **102:** 1-21, 1996. Crabs 자료 파일의 원본은 본문 웹사이트에 있다.

[2] 설명과 그림 자료는 https://en.wikipedia.org/wiki/Horseshoe_crab에서 확인하라.
[3] www.bluffton.com/wp-content/uploads/Spawning-horseshoe-crabs.jpg를 확인하라.

이 표본에서 등딱지 너비는 평균이 26.3 cm이고 표준편차는 2.1 cm이다. 〈표 3.2〉는 173마리의 암참게 중 12마리의 암참게에 대한 너비를 포함한 총 4개 설명변수들의 자료이다.

〈그림 3.3〉은 너비와 반응변수인 부수체 도수에 대한 산점도인데 각 점에 있는 수치는 그 점에서의 반응 도수값을 나타내고 있다. 도수값들이 상당히 넓게 분포되어 있어서 명확한 추세를 확인하기 어렵다. 몇몇 소프트웨어들을 이용하면 자료를 평활시켜(smoothing) 전반적인 추세를 볼 수 있는 좀 더 정교한 방법을 사용할 수 있다. 11.4절에서 소개될 **일반적인 가법모형**(generalized additive model)에 기초한 평활법은 GLM보다 더 일반적인 구조 형태를 제공해 준다. 이러한 방법은 일정한 형태를 가지면서 가장 좋은 예측을 제공해주는 가능한 설명변수들의 복잡한 함수관계를 찾아낸다. 〈그림 3.3〉은 이러한 방법을 사용하여 평활된 곡선을 보여 주고 있다. 평활된 곡선은 강한 증가 추세를 보이고 있다. 이 그림에서 근사적으로 너비에 대한 선형추세를 볼 수 있으므로 이제 너비의 평균이나 로그 평균이 너비의 선형식으로 표현되는 모형에 대해서 논의해 보자.

μ를 암참게 한 마리가 갖고 있는 부수체 수의 기댓값이라고 하고 x를 암참게의 너비라고 하자. GLM 소프트웨어로부터 포아송 로그 선형모형을 ML 방법으로 적합한 결과는 다음과 같다.

$$\log\hat{\mu} = \hat{\alpha} + \hat{\beta}x = -3.305 + 0.164x$$

$\hat{\beta} > 0$이므로 너비는 부수체 수에 대해 양의 효과를 가지고 있음을 알 수 있다. 다음은 R에서 모형을 적합시키기 위한 glm 함수를 사용하는 프로그램이다.

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> Crabs
    crab  sat  weight  width   color spine  # sat = number of satellites
1      1    8   3.050   28.3       2     3  # showing 3 of 173 observations
2      2    0   1.550   22.5       3     3
...
173  173    0   2.000   24.5       2     2
> plot(sat ~ width, xlab="Width", ylab="Number of satellites", data=Crabs)
> fit <- glm(sat ~ width, family=poisson(link=log), data=Crabs)
> # canonical link for Poisson is log, so "(link=log)" is not necessary
> summary(fit)
             Estimate    Std. Error
(Intercept)  -3.30476       0.54224
width         0.16405       0.01997
> library(gam) # generalized additive model smoothing fit
> gam.fit <- gam(sat ~ s(width), family=poisson, data=Crabs)
> # s() is smooth function predictor for generalized additive model
> curve(predict(gam.fit, data.frame(width=x), type="resp"), add=TRUE)
------------------------------------------------------------------------------
```

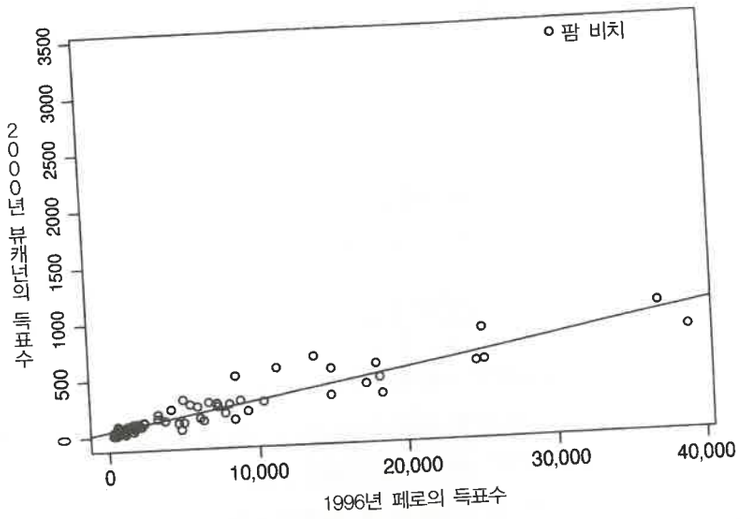

그림 3.6 구역에서 1996년 페로 후보와 2000년도 개혁당 후보 뷰캐넌의 전체 득표수

3.5 웹사이트 www.stat.ufl.edu/~aa/cat/data로부터 〈표 3.2〉에 있는 참게 자료를 다운받을 수 있다. 만일 참게가 적어도 한 마리의 부수체를 가지면 $y=1$이라 하고 그렇지 않으면 $y=0$이라 하자.

a. 무게를 예측변수로 이용해서 선형확률모형을 $P(Y=1)$에 적합하라. 만약 소프트웨어가 이항분포에 대해서 항등연결함수를 사용할 수 없거나 수렴하지 않으면 Y를 정규분포를 따르는 것처럼 간주하여 보통의 최소제곱법을 이용하여 모수를 추정하고 추정값들을 해석하라. 가장 큰 무게값 5.20 kg에 대하여 $P(Y=1)$을 예측하고 결론을 내려라.

b. Y를 이항변수로 간주하고 로지스틱 모형을 적합하라. 무게값 5.20 kg에 대하여 $\hat{P}(Y=1)=0.9968$이 됨을 보여라.

3.6 2016년도 일반사회조사에서 18세에서 27세까지를 대상으로 정치성향(1=매우 보수적, 2=보수적, 3=약간 보수적, 4=중간, 5=약간 진보적, 6=진보적, 7=매우 진보적)과 정당(민주당, 공화당)에 대하여 교차분할표를 구한 것이다.

```
-----------------------------------------------------
             1  2  3  4  5  6  7
Democrat     5 18 19 25  7  7  2
Republican   1  3  1 11 10 11  1
-----------------------------------------------------
```

지지정당이 민주당일 확률에 미치는 정치성향의 효과를 보기 위해 R을 이용하

여 모형을 적합한 결과가 다음과 같다.

```
-------------------------------------------------------------
> y <- c(5,18,19,25,7,7,2); n <- c(6,21,20,36,17,18,3)
> x <- c(1,2,3,4,5,6,7)
> fit <- glm(y/n ~ x, family=binomial(link=logit), weights=n)
> summary(fit)
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)  3.1870 0       .7002    4.552   5.33e-06
x             -0.5901      0.1564   -3.772   0.000162
---
    Null deviance: 24.7983 on 6 degrees of freedom
Residual deviance:  7.7894 on 5 degrees of freedom
Number of Fisher Scoring iterations: 4
> confint(fit)
                2.5 %      97.5 %
(Intercept)   1.90180     4.66484
x            -0.91587    -0.29832
-------------------------------------------------------------
```

a. 예측식을 제시하고 정치성향 효과의 방향에 대해서 해석하라.

b. 정치성향 효과에 대한 95% 왈드 신뢰구간을 구하고 위에 주어진 프로파일 가능도 신뢰구간과 비교하라.

c. 정치성향 효과에 대한 왈드 검정을 수행하라. 검정통계량값과 P-값을 제시하고 결과를 해석하라.

d. 가능도비 검정을 수행하고 검정통계량값과 P-값을 제시하고 결과를 해석하라.

e. R 출력 결과에 있는 Fisher 스코어 알고리즘의 반복수에 대해서 설명하라.

3.7 〈표 3.1〉의 코골이와 심장병 자료에 대하여

a. (i) (0, 2, 4, 6), (ii) (0, 1, 2, 3), (iii) (1, 2, 3, 4)의 점수들을 이용하여 로지스틱 회귀모형을 다시 적합하라. 이 세 점수들에 대한 모수추정값을 서로 비교하라. 세 가지 모형의 모수추정값들을 비교하라. 적합된 값들을 비교하라. 점수들 간격의 상대적 크기를 유지하는 점수들의 선형 변환 효과에 대해서 어떤 결론을 내릴 수 있는가?

b. 일주일 동안 코고는 날짜의 빈도를 반영하는 (0, 2, 6, 7) 점수를 사용하여 로지스틱 회귀모형을 적합하라. 이 책에서 사용한 (0, 2, 4, 5) 점수의 적합 결과와 비교하라. 이 결과가 점수의 선택에 민감하다고 할 수 있는가?

3.8 3.2.3절의 〈표 3.1〉의 코골이와 심장병 자료에 대하여 로지스틱 회귀모형을 적합하라. 코골이 효과에 대한 유의성검정 결과와 신뢰구간을 제시하라.

3.9 〈표 3.4〉의 Credit 자료 파일은 이 책의 웹사이트에서 다운로드한 것이다. 연수입과 소유하고 있는 여행 신용카드(American Express, Diners Club과 같은) 수와의 관계를 알기 위해 임의로 선택한 100명의 이탈리아인들로부터 구한 표본자료이다. 아래의 표는 연수입(단위 천 유로)을 나타내는 x의 각 추출된 사람의 수와 이 중에서 한 개 이상의 여행 신용카드를 소유한 사람의 수를 나타내었다. 로지스틱 회귀모형을 사용하여 소프트웨어로 분석한 결과는 다음과 같다.

```
---------------------------------------------
             Estimate   Std. Error
(Intercept)   -3.5179       0.7103
x              0.1054       0.0262
---------------------------------------------
```

a. 예측식을 구하고 $\hat{\beta}$의 부호를 해석하라.

b. $\hat{P}(Y=1)=0.50$일 때, 추정된 로짓값이 0임을 보여라. 이것을 기초로 하여 이 자료에서 연수입이 33.4천 유로일 때 여행 신용카드를 소지할 예측확률이 0.50인 이유를 설명하라.

c. 소프트웨어를 이용하여 이 책의 웹사이트에서 Credit 자료 파일을 불러와서 어떻게 로지스틱 회귀모형적합 결과를 얻을 수 있는지 설명하라.

3.10 최근의 일반사회조사 중에서 "당신의 직장에서 몇 명의 사람들이 친한 친구인가?"라는 질문이 있다. 756명의 응답자 중에서 평균이 2.76명이고 표준편차가 3.65이고 최빈값은 0이였다. 포아송 분포가 이 자료를 잘 설명하는지 여부를 답하고 그 이유를 설명하라.

▸표 3.4 연습문제 3.9의 이탈리아인들의 신용카드에 대한 자료

수입	사례의 수	신용 카드 수	수입	사례의 수	신용 카드 수	수입	사례의 수	신용 카드 수
12	1	0	21	2	0	34	3	3
13	1	0	22	1	1	35	5	3
14	8	2	24	2	0	39	1	0
15	14	2	25	10	2	40	1	0
16	9	0	26	1	0	42	1	0
17	8	2	29	1	0	47	1	0
19	5	1	30	5	2	60	6	6
20	7	0	32	6	6	65	1	1

출처: 이탈리아 Milan에 있는 Bocconi 대학의 교수 R. Piccarreta.

3.11 컴퓨터 칩의 실리콘 기판을 제조하는 데 사용되는 두 공정 과정의 결함율을 분석하기 위해 실험을 하였다. 10개의 기판에 대해 처리 A를 적용하였을 때에 기판에서 관측된 결함 수는 각각 8, 7, 6, 6, 3, 4, 7, 2, 3, 4이었다. 10개의 기판에 대하여 처리 B를 적용하였을 때에는 결함 수가 9, 9, 8, 14, 8, 13, 11, 5, 7, 6이었다. 결함 수를 각각 평균 μ_A와 μ_B를 갖는 서로 독립인 포아송 변량이라고 간주하자. 모형 $\log\mu=\alpha+\beta x$을 고려하자. 여기서 처리 A는 $x=0$, 처리 B는 $x=1$로 나타낸다. 그러면 $\beta=\log\mu_B-\log\mu_A=\log(\mu_B/\mu_A)$이고 $e^{\beta}=\mu_B/\mu_A$가 된다. 이 모형을 적합하고 예측식을 제시하고 $\hat{\beta}$값을 해석하라.

3.12 연습문제 3.11을 참조하라.

a. $H_0:\mu_A=\mu_B$에 대한 검정을 $H_0:\beta=0$의 왈드 검정이나 가능도비 검정을 통하여 실시하라.

b. μ_B/μ_A에 대한 95% 신뢰구간을 구하라. (힌트: 우선 $\beta=\log\mu_B-\log\mu_A=\log(\mu_B/\mu_A)$에 대한 신뢰구간을 구하라.)

3.13 〈표 3.2〉의 참게 예제 자료를 www.stat.ufl.edu/~aa/cat/data 웹사이트에서 다운로드하여 x=무게를 예측변수로 간주하고 Y=부수체의 수를 반응변수로 간주하여 포아송 로그 선형모형을 적합하라.

a. 예측식을 구하고 암참게의 평균 무게가 2.44 kg일 때 부수체의 평균수를 추정하라.

c. $\hat{\beta}$를 이용하여 무게의 효과를 설명하라. 모수 β와 무게가 1 kg 증가할 때마다 발생하는 승법 효과에 대한 95% 신뢰구간을 구하라.

d. 부수체의 평균수와 무게와의 독립성을 왈드 검정과 가능도비 검정을 하고 결과를 해석하라.

3.14 도수자료를 모형화할 때에 왜 보통의 회귀분석모형에서 쓰는 원잔차 $(y_i-\hat{\mu}_i)$를 사용하는 것이 충분하지 않은지 설명하라.

3.15 다음 물음에 대하여 참인지 거짓인지를 판단하라.

a. Y가 정규분포를 따르는 것으로 간주하는 회귀분석모형은 정규 랜덤성분과 항등연결함수를 갖는 GLM의 특별한 경우이다.

a. GLM에서 Y는 꼭 정규분포를 따르지 않아도 된다. GLM은 Y의 평균값