Logistic Regression

2024-06-18

Example 4.1.1

$$y = \begin{cases} 1 & \text{at least one satellite} \\ 0 & \text{no satellite} \end{cases} \tag{1}$$

```r
library(haven)
```

```
## Warning: package 'haven' was built under R version 4.3.3
```

```r
crab= read.csv("Crabs1.dat", sep="")
attach(crab)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     crab
```

```r
head(crab)
```

```
##   crab sat y weight width color spine
## 1    1   8 1   3.05  28.3     2     3
## 2    2   0 0   1.55  22.5     3     3
## 3    3   9 1   2.30  26.0     1     1
## 4    4   0 0   2.10  24.8     3     3
## 5    5   4 1   2.60  26.0     3     3
## 6    6   0 0   2.10  23.8     2     3
```

```r
#View(crab)
summary(crab)
```

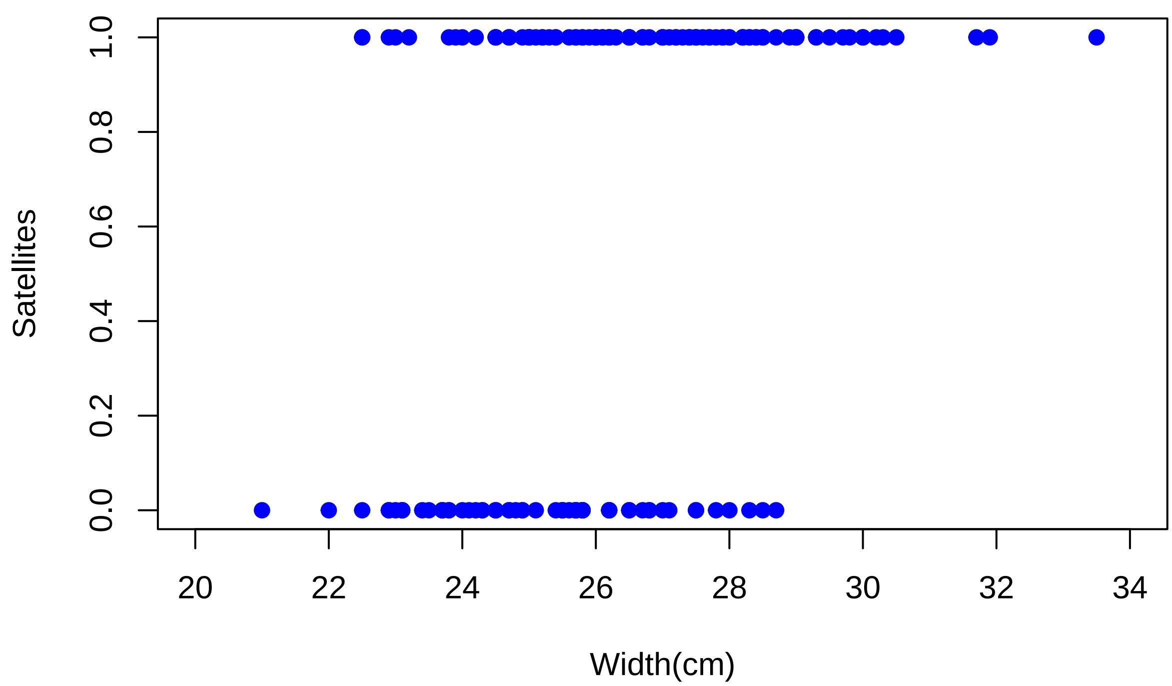
```
##       crab          sat               y              weight          width
##  Min.   :  1   Min.   : 0.000   Min.   :0.0000   Min.   :1.200   Min.   :21.0
##  1st Qu.: 44   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:24.9
##  Median : 87   Median : 2.000   Median :1.0000   Median :2.350   Median :26.1
##  Mean   : 87   Mean   : 2.919   Mean   :0.6416   Mean   :2.437   Mean   :26.3
##  3rd Qu.:130   3rd Qu.: 5.000   3rd Qu.:1.0000   3rd Qu.:2.850   3rd Qu.:27.7
##  Max.   :173   Max.   :15.000   Max.   :1.0000   Max.   :5.200   Max.   :33.5
##      color           spine
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000
##  Mean   :2.439   Mean   :2.486
##  3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :3.000
```

```r
crab$z= ifelse(crab$sat>=1, 1,0)
head(crab)
```

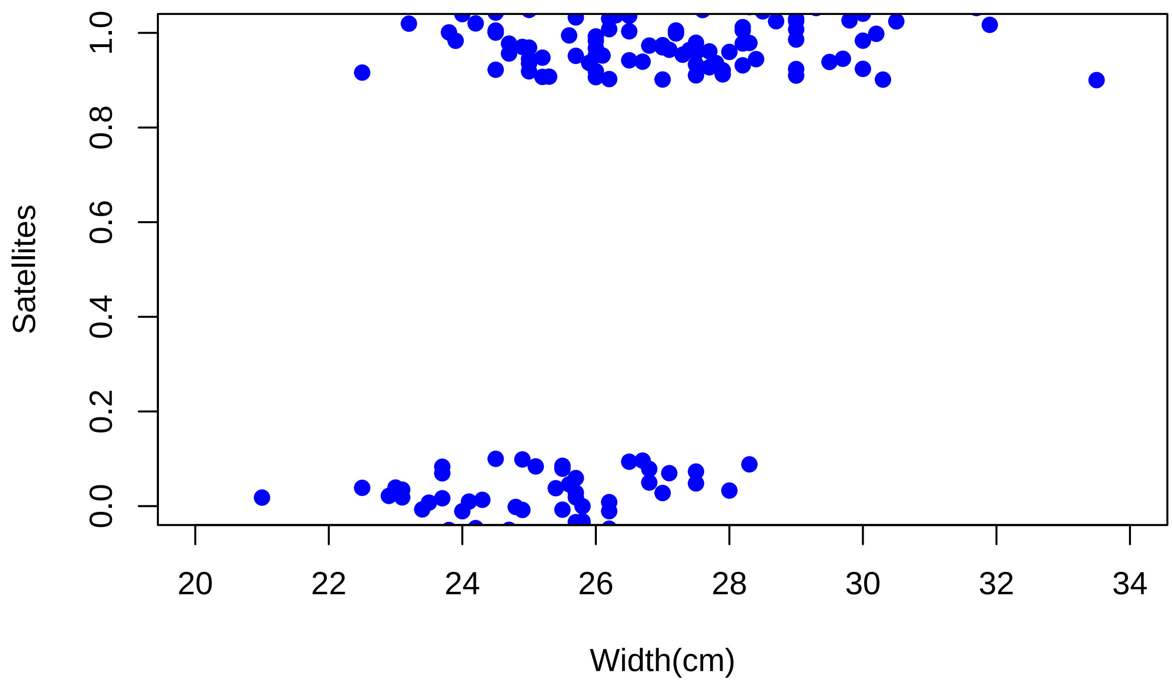
```
##   crab sat y weight width color spine z
## 1    1   8 1   3.05  28.3     2     3 1
## 2    2   0 0   1.55  22.5     3     3 0
## 3    3   9 1   2.30  26.0     1     1 1
## 4    4   0 0   2.10  24.8     3     3 0
## 5    5   4 1   2.60  26.0     3     3 1
## 6    6   0 0   2.10  23.8     2     3 0
```

Scatter plot

```r
plot(x=width, y=y, ylab="Satellites", xlab="Width(cm)", col="blue", pch=19, xlim=c(20,34), ylim=c(0,1))
```




```r
plot(jitter(y,0.5)~width, data=crab, ylab="Satellites", xlab="Width(cm)", col="blue", pch=19, xlim=c(20
```



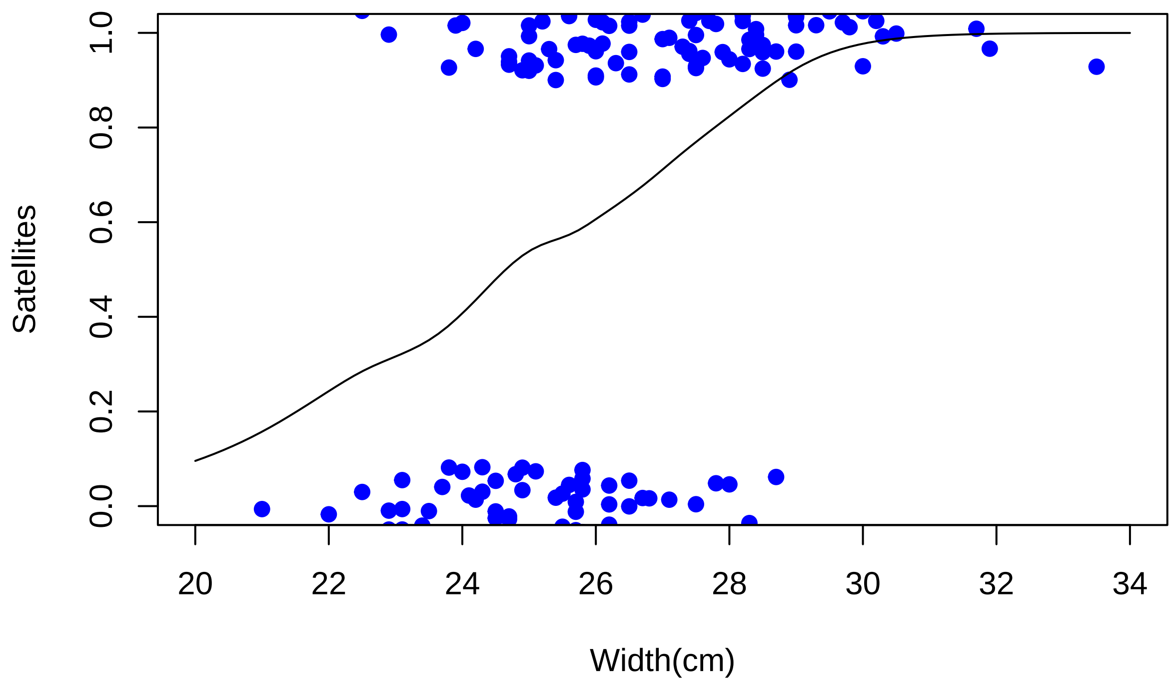
Generalized additive models

```
#install.packages("gam")
library(gam)
```

```
## Warning: package 'gam' was built under R version 4.3.3

## Loading required package: splines

## Loading required package: foreach

## Warning: package 'foreach' was built under R version 4.3.3

## Loaded gam 1.22-3
```

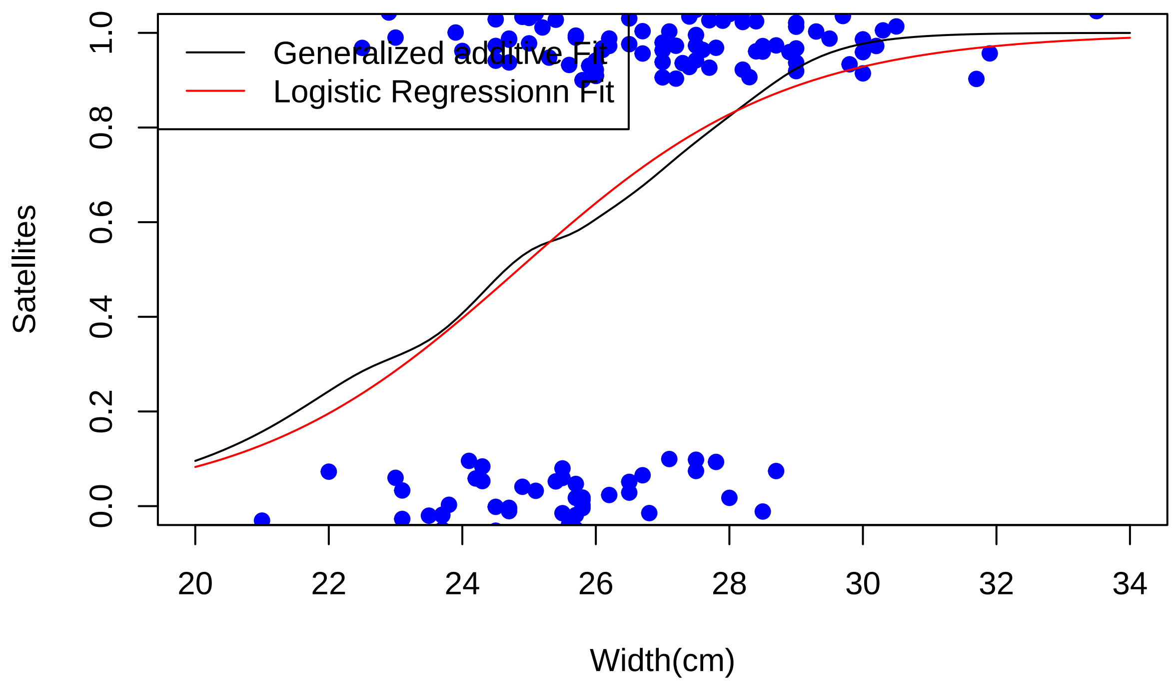
```
gam.fit= gam(y~s(width), family=binomial, data=crab) #s=smooth function
```

```
plot(jitter(y,0.5)~width, data=crab, ylab="Satellites", xlab="Width(cm)", col="blue", pch=19, xlim=c(20
curve(predict(gam.fit,data.frame(width=x),type="resp"), add=TRUE, col="black")
```



Logistic Regression where x=width

```
plot(jitter(y,0.5)~width, data=crab, ylab="Satellites", xlab="Width(cm)", col="blue", pch=19, xlim=c(20
curve(predict(gam.fit,data.frame(width=x),type="resp"), add=TRUE, col="black")

fit= glm(y~width, family=binomial, data=crab)
curve(predict(fit, data.frame(width=x),type="resp"), add=TRUE, col="red")

legend("topleft",c("Generalized additive Fit", "Logistic Regressionn Fit"), lty=c(1,1), col=c("black","
```



#Finding the minimum width

```
which(crab$width==min(width))
```

```
## [1] 14
```

```
crab$y[14]
```

```
## [1] 0
```

#Fit the model

```
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ width, family = binomial, data = crab)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
## width         0.4972     0.1017   4.887 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 194.45  on 171  degrees of freedom
## AIC: 198.45
##
## Number of Fisher Scoring iterations: 4
```

```r
predict(fit, data.frame(width=c(min(crab$width),max(crab$width))), type="response")
```

```
##         1         2
## 0.1290960 0.9866974
```

```r
predict(fit, data.frame(width= mean(crab$width)), type="response")
```

```
##         1
## 0.6738768
```

```r
#Median effective level

fit$coefficients[1]/fit$coefficients[2]
```

```
## (Intercept)
##   -24.83922
```

```r
#when width increases by 1cm the odds of having 1 satellite is found by exp(beta)
exp(fit$coefficients[2])
```

```
##    width
## 1.644162
```

Profile likelihood CI

```r
P=confint(fit)
```

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

```r
P
```

```
##                   2.5 %     97.5 %
## (Intercept) -17.8100090 -7.4572470
## width         0.3083806  0.7090167
```

```r
exp(P[2,])
```

```
##    2.5 %   97.5 %
## 1.361219 2.031992
```

We infer that a 1-cm increase in width has at least a 36% increase and at most a doubling in the odds that a female crab has a satellite

Wald CI

```r
W=confint.default(fit)
W
```

```
##                  2.5 %     97.5 %
## (Intercept) -17.5030100 -7.1986254
## width         0.2978326  0.6966286
```

```r
exp(W[2,])
```

```
##    2.5 %   97.5 %
## 1.346936 2.006975
```

#Estimate when prob of satellite x=26.5

```r
New=data.frame(width=26.5)
Fitted=predict(fit, New, type="response")
Fitted
```

```
##         1
## 0.6954646
```

```r
fit$coefficients[1]+fit$coefficients[2]*26.5
```

```
## (Intercept)
##   0.8257928
```

95% confident interval for estimated prob

#Fit values and CI for probabilities Step1: Construct a 95% CI for P(Y=1) at x=26.5

Linear predictor= $\hat{\alpha} + \hat{\beta} * 26.5$

```r
lp=predict(fit, New, se.fit=TRUE)
lp
```

```
## $fit
##         1
## 0.8257928
##
## $se.fit
## [1] 0.1886957
##
## $residual.scale
## [1] 1
```

```r
alpha=0.05
a=1-(alpha/2)
z=qnorm(a)
z
```

```
## [1] 1.959964
```

```r
lp$fit +c(-1,1)*z*lp$se.fit #CI for linear predictor
```

```
## [1] 0.455956 1.195630
```

Step 2: Applying the $exp()/[1 + exp()]$ transform to endpoints

Confidence bounds for $\pi_{x_0} = P(Y = 1)$

```r
CI=exp(lp$fit+c(-1,1)*z*lp$se.fit)/(1+exp(lp$fit+c(-1,1)*z*lp$se.fit))
CI
```

```
## [1] 0.6120544 0.7677464
```

```r
#Fitted value and CI for P(Y=1)
cbind(Fitted,"CI:LB"=CI[1],"CI:UB"=CI[2])
```

```
##      Fitted     CI:LB     CI:UB
## 1 0.6954646 0.6120544 0.7677464
```

For female crabs of width x = 26.5, which is near the mean width, the estimated probability of a satellite is P(Yd= 1) = 0.695 and a 95% confidence interval for P(Y = 1) is (0.61, 0.77).

Graph Model to Estimate Probabilities with CI

The following graph shows the estimated probabilities and the lower and upper 95% confidence bands.

Sample Proportion

```r
total=length(crab$y)
total #total crabs
```

```
## [1] 173
```

```r
#how many crabs have width=26.5cm
n=length(which(crab$width==26.5))
n
```

```
## [1] 6
```

```r
x=length(which(crab$width==26.5 & crab$y==1))
x
```

```
## [1] 4
```

```r
phat=x/n
phat
```

```
## [1] 0.6666667
```

```r
SE= sqrt(phat*(1-phat)/n)
SE
```

```
## [1] 0.1924501
```

Model based estimate

```r
predict(fit, New,  type="response", se.fit=TRUE)
```

```
## $fit
##         1
## 0.6954646
##
## $se.fit
##          1
## 0.03996454
##
## $residual.scale
## [1] 1
```

Question: Construct a 95% CI for P(Y=1) at x=25cm

Linear predictor= $\hat{\alpha} + \hat{\beta} * 25$

x= weight

```r
fit1= glm(y~weight, family=binomial, data=crab)
summary(fit1)
```

```
##
## Call:
## glm(formula = y ~ weight, family = binomial, data = crab)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
## weight        1.8151     0.3767   4.819 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 195.74  on 171  degrees of freedom
## AIC: 199.74
##
## Number of Fisher Scoring iterations: 4
```

Logistic Regression Equation is:

logit(phat(x))= -0.14487+ 0.3227x

```r
m=mean(crab$weight)
m
```

```
## [1] 2.437191
```

```r
#vector of values
A=predict(fit1, data.frame(weight=c(min(weight),max(weight))), type="response")


B=predict(fit1, data.frame(weight=m), type="response")

cbind("min"=A[1], "max"=A[2], "mean"=B[1])
```

```
##         min       max      mean
## 1 0.1799697 0.9968084 0.6746137
```

At the minimum weight in this sample of 1.2 kg, the estimated probability that the crab has at least one satellite is 0.242373.

At the maximum sample weight of 5.2 kg, the estimated probability equals 1.533186.

#Calculate the median effective level

```r
itr=fit1$coefficients[1]
itr
```

```
## (Intercept)
##   -3.694726
```

```r
w=fit1$coefficients[2]
w
```

```
##   weight
## 1.815145
```

```r
m= itr/w
m
```

```
## (Intercept)
##     -2.0355
```

Calculate the incremental rate of change in the fitted probability at the sample mean and interpret the value.

```r
p= fit1$coefficients[2]*B[1]*(1-B[1])
p
```

```
##    weight
## 0.3984425
```

For female crabs near the mean weight, the estimated probability of having at least one satellite increases at the rate of 0.07420375 per 1-kg increase in weight.

Profile Likelihood CI

```r
confint(fit1)
```

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

```
##                 2.5 %    97.5 %
## (Intercept) -5.505932 -2.039701
## weight       1.113790  2.597305
```

We infer that a 1-kg increase in weight has at least a 11% increase and at most 59% increase in the odds that a female crab has a satellite.

Wald CI

```r
confint.default(fit1)
```

```
##                 2.5 %    97.5 %
## (Intercept) -5.419882 -1.969571
## weight       1.076834  2.553455
```