Introduction to Logistic Regression

GLMs for Binary Response Data

Example: Crab species identification

We will work with a data set about Leptograpsus crabs originally presented in

Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22, 417–425.

They have also been discussed previously in

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

There are two species of this crab; we will examine the effect of certain physical dimenions on crab species. The data we are working with contains 5 morphological measurements on 200 crabs, 100 each of two species of Leptograpsus crabs collected at Fremantle, W. Australia.

The variables in this data set are as follows:

- `sp`: species - "B" or "O" for blue or orange.
- `sex`: the crab's sex
- `FL`: frontal lobe size (mm).
- `RW`: rear width (mm).
- `CL`: carapace length (mm).
- `CW`: carapace width (mm).
- `BD`: body depth (mm).

For purposes of this example, we will only focus on species and frontal lobe size.

Binary encoding of response variable

Typically in logistic regression, we use an indicator variable for the response variable:

$$Y_i = \begin{cases} 1 & \text{if crab number } i \text{ is an orange crab} \\ 0 & \text{otherwise (if a blue crab)} \end{cases}$$

```
crabs <- crabs %>%
  mutate(
    sp_01 = ifelse(sp == "O", 1, 0)
  )
head(crabs)
```

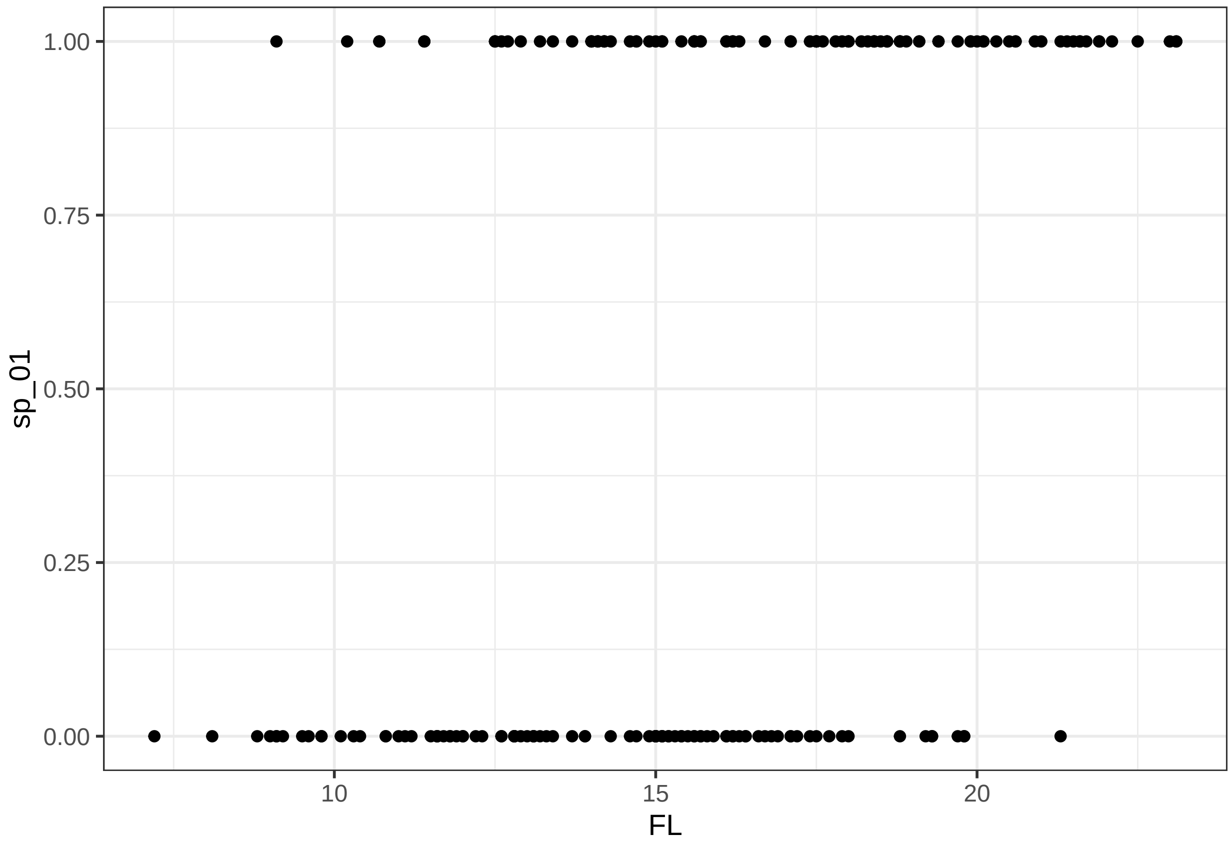
```
##   sp sex   FL   RW   CL   CW   BD sp_01
## 1  O   F 21.4 18.0 41.2 46.2 18.7     1
## 2  O   M 15.1 11.4 30.2 33.3 14.0     1
## 3  O   M 18.8 13.4 37.2 41.1 17.5     1
## 4  O   F 22.5 17.2 43.0 48.7 19.8     1
## 5  O   M 14.2 10.7 27.8 30.9 12.7     1
## 6  B   M 17.9 14.1 39.7 44.6 16.8     0
```

```
dim(crabs)
```

```
## [1] 200   8
```

Plot of the data

```
ggplot(data = crabs, mapping = aes(x = FL, y = sp_01)) +
  geom_point() +
  theme_bw()
```



Fit logistic regression model

Note:

- Behind the scenes, sp is converted to 0/1 representation by the glm function
- By default, assignment is in alphabetic order, so "B" goes to 0 and "O" goes to 1.

```
logistic1 <- glm(sp_01 ~ FL, data=crabs, family = binomial)
```

Print model summary

```
summary(logistic1)
```

```
##
## Call:
## glm(formula = sp_01 ~ FL, family = binomial, data = crabs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93977  -1.02847   0.02585   0.94443   2.03821
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.67278    0.83671  -5.585 2.34e-08 ***
## FL           0.29994    0.05278   5.683 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.26  on 199  degrees of freedom
## Residual deviance: 235.48  on 198  degrees of freedom
## AIC: 239.48
##
## Number of Fisher Scoring iterations: 3
```

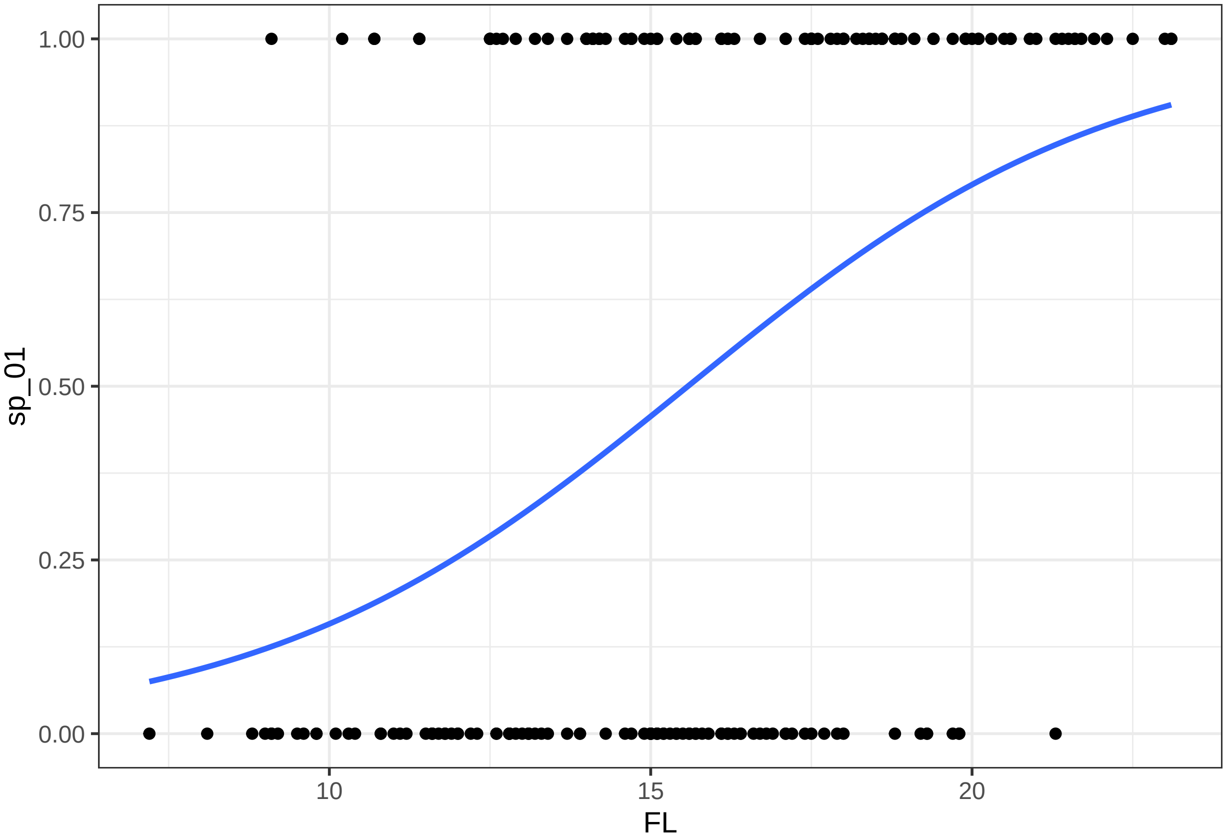
```
## Effect of FL on species odds
exp(logistic1$coefficients[2])
```

```
##       FL
## 1.349784
```

Plot of the data with logistic regression model fit

This curve represents the estimated probability that a crab is orange, as a function of frontal lobe size (mm).

```
ggplot(data = crabs, mapping = aes(x = FL, y = sp_01)) +
  geom_point() +
  geom_smooth(method="glm", method.args=list(family="binomial"), se=FALSE) +
  theme_bw()
```



1. Based on this model, how could you calculate the estimated probability that a crab with a frontal lobe size of 15 mm is orange?

Method 1

```
b_hat <- matrix(logistic1$coefficients)
X <- matrix(c(1,15), nrow=1)

exp(X%*%b_hat)/(1+exp(X%*%b_hat))
```

```
##           [,1]
## [1,] 0.4567046
```

Method 2

```
predict_data <- data.frame(
  FL = 15
)
predict(logistic1, newdata=predict_data, type="response")
```

```
##         1
## 0.4567046
```

2. What is the interpretation of $\hat{\beta}_1$ in terms of odds?

3. What is the estimated relationship between the odds that a crab with a frontal lobe of 10 mm is orange versus a crab with a frontal lobe of 20 mm is orange?

```
X <- cbind(c(1,1),
           c(10,20))
exp(X[2,]%*%b_hat)/exp(X[1,]%*%b_hat)
```

```
##          [,1]
## [1,] 20.07437
```

```
exp(b_hat[2]*10)
```

```
## [1] 20.07437
```

4. Mathematical check - why are these both valid ways to calculate the estimated relationship?