

# Stat 340: Intro. to Classification and Logistic Regression

## 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 attempt to predict the species of a crab based on measurements of its physical dimensions. 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).

### Binary encoding of response variable

It is traditional in logistic regression to 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) # 1 if orange, 0 otherwise
  )
head(crabs)
```

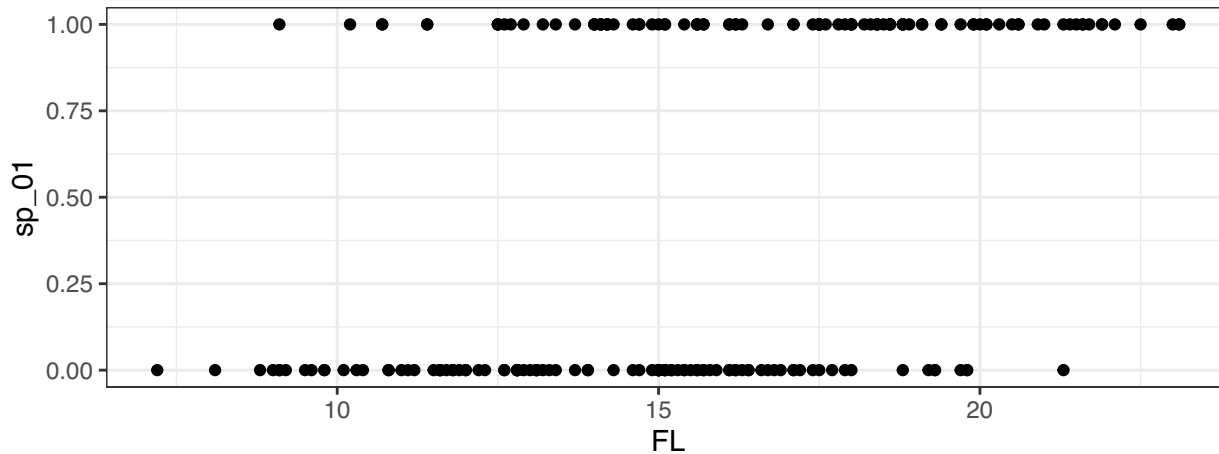
```
##      sp sex  FL  RW  CL  CW  BD sp_01
## 196  O   F 21.4 18.0 41.2 46.2 18.7     1
## 120  O   M 15.1 11.4 30.2 33.3 14.0     1
## 137  O   M 18.8 13.4 37.2 41.1 17.5     1
## 199  O   F 22.5 17.2 43.0 48.7 19.8     1
## 116  O   M 14.2 10.7 27.8 30.9 12.7     1
## 42   B   M 17.9 14.1 39.7 44.6 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()
```



## Train logistic regression model

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

```
logistic_fit <- train(  
  form = sp ~ FL,  
  data = crabs,  
  family = "binomial", # this is an argument to glm; response is 0 or 1, binomial  
  method = "glm",      # method for fit; "generalized linear model"  
  trControl = trainControl(method = "none")  
)
```

## Print model summary (coefficient estimates and p-values for tests)

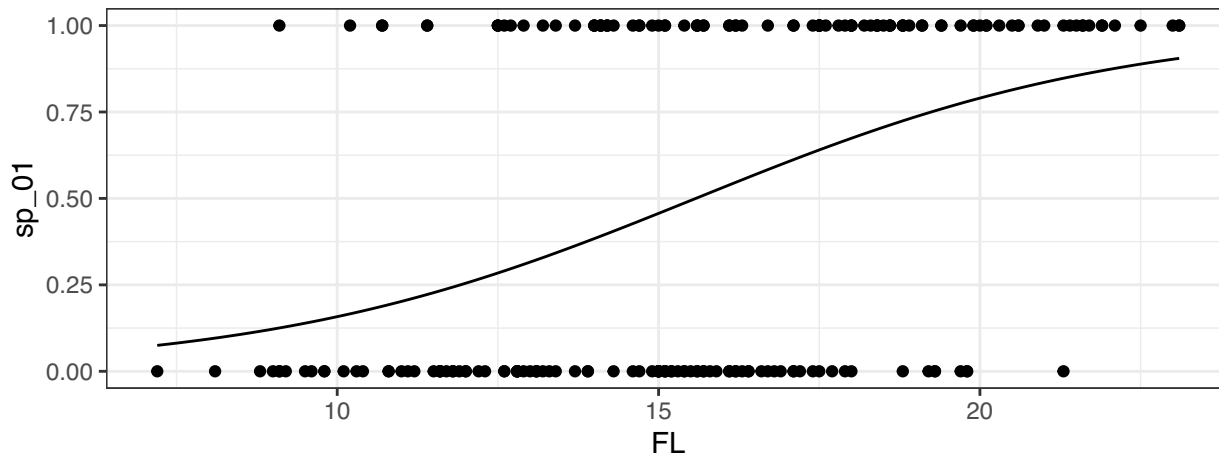
```
summary(logistic_fit$finalModel)
```

```
##  
## Call:  
## NULL  
##  
## 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
```

## Plot estimated class probability function

Note: code is essentially identical to what we used for KNN.

```
predict_logistic_probs <- function(x, sp) {  
  f_hats <- predict(logistic_fit, newdata = data.frame(FL = x), type = "prob")  
  f_hats[[sp]]  
}  
ggplot(data = crabs, mapping = aes(x = FL, y = sp_01)) +  
  geom_point() +  
  stat_function(fun = predict_logistic_probs, args = list(sp = "0")) +  
  ylim(0, 1) +  
  theme_bw()
```



## What's the interpretation of $\hat{\beta}_1$ ?

For an increase of 1mm in the size of the frontal lobe of a crab, we expect the log odds of the crab to be orange to increase by 0.29994.

For an increase of 1mm in the size of the frontal lobe of a crab, we expect the odds of the crab to be orange to increase by a factor of  $\exp(0.29994) = 1.3498$ .

What's the estimated probability that a crab with a frontal lobe size of 20 mm is an orange crab? Calculate by hand too.

```
predict(logistic_fit, newdata = data.frame(FL = 20), type = "prob")
```

```
##           B           0  
## 1 0.2098043 0.7901957
```

$$\hat{Pr}(Y = 1|FL = 20) = \frac{e^{-4.6728+0.29994 \cdot 20}}{1 + e^{-4.6728+0.29994 \cdot 20}} = 0.7902$$

What's the predicted species for a crab with a frontal lobe size of 20 mm?

```
predict(logistic_fit, newdata = data.frame(FL = 20))
```

```
## [1] 0  
## Levels: B 0
```

Orange, since the probability of being an orange crab =  $0.7802 > 0.5$ .

Code suppressed, but here's a comparison to a KNN fit with  $K = 30$  neighbors:

