Stat 230 - Lab 11 - Logistic Regression - Solution

P.B. Matheson adapted from A.S. Wagaman

This lab will lead you through some logistic regression examples. Some code is provided. In some cases you will be providing your own code. As we have not yet seen how to compare models (no comparable quantity like R-squared, for example), the focus of this lab is on fitting and assessing models with one predictor. We start off just fitting models and then moving to assessing them.

Empirical Logit Plot Function

Again, we have a function to create empirical logit plots. It requires a numeric 0/1 response. If you have a factor as the response, convert it to numeric for this to run.

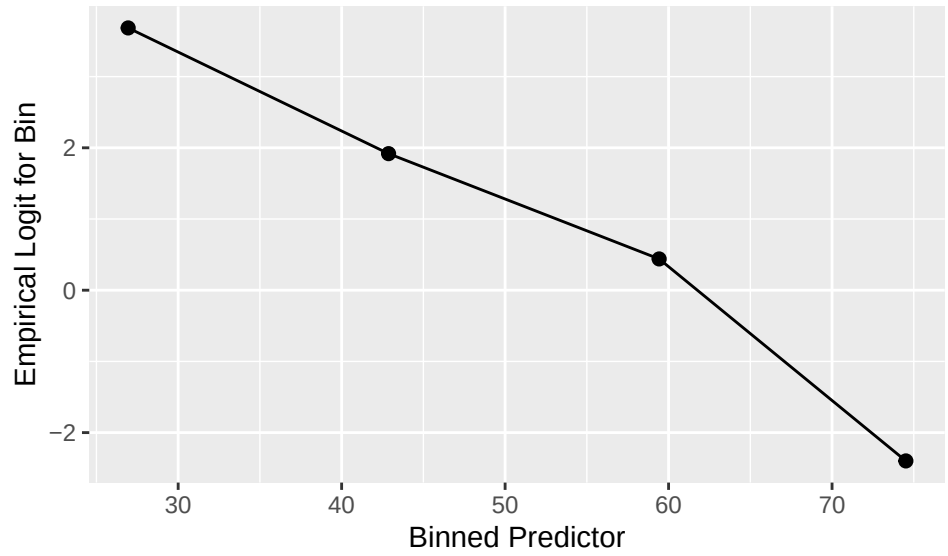
```
emplogitplot <- function(resp, pred, numbreak = 10) {
  # assumes resp is dichotomous with values 0 and 1
  tmpGroup <- cut(pred, breaks = numbreak)
  binned.y <- mosaic::mean(~ resp | tmpGroup)
  binned.x <- mosaic::mean(~ pred | tmpGroup)
  logy <- mosaic::logit(binned.y)
  ds <- data.frame(logy, binned.x)
  gf_point(logy ~ binned.x, cex = 2, pch = 19, data=ds) %>%
    gf_line() %>%
    gf_labs(x = "Binned Predictor", y = "Empirical Logit for Bin")
}

#call this as:
#with(data set, emplogitplot(responsevariablename, predictorvariablename, numbreaks/bins))
```

Here is a quick example:

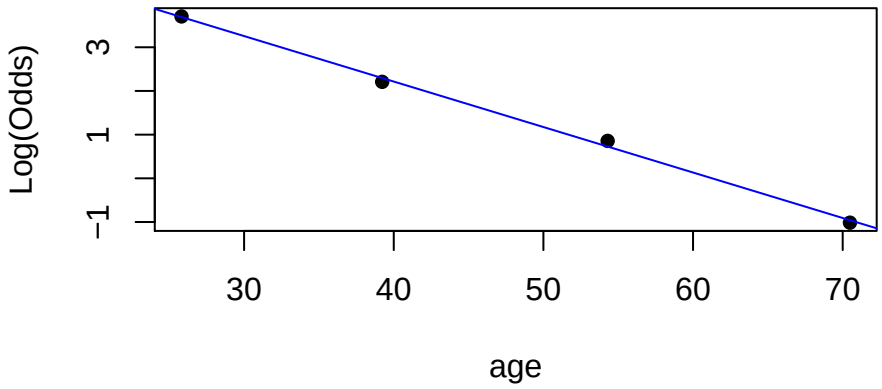
```
data(Whickham)
Whickham <- mutate(Whickham, isAlive = 2 - as.numeric(outcome)) #create numeric version of response 0/1

with(Whickham, emplogitplot(isAlive, age, 4))
```



The textbook provides an alternative function that you can use if you like how it works and shows the resulting output. It fits a line so you can see if the points follow it, rather than connecting the points with lines. To use this function is slightly different syntax:

```
emplogitplot1(isAlive~age,ngroups=4,data=Whickham)
```



If you want to use either of these functions in a different .Rmd, remember for the first one, you need to copy over the function definition (the chunk that defines it), or for the second one, you have to be sure to load the Stat2Data package.

Logistic Regression Examples

```
fly <- read.table("https://pmatheson.people.amherst.edu/stat230/bitingfly.txt", header = T, sep = "")
glimpse(fly)
```

Example 1

```
## Rows: 70
## Columns: 8
## $ wingl    <int> 85, 87, 94, 92, 96, 91, 90, 92, 91, 87, 97, 89, 94, 96, 10...
## $ wingw    <int> 41, 38, 44, 43, 43, 44, 42, 43, 41, 38, 45, 38, 45, 44, 49...
## $ thirdpl  <int> 31, 32, 36, 32, 35, 36, 36, 36, 36, 35, 39, 36, 37, 37, 35...
## $ thirdpw  <int> 13, 14, 15, 17, 14, 12, 16, 17, 14, 11, 17, 13, 13, 14, 14...
## $ fourthpl <int> 25, 22, 27, 28, 26, 24, 26, 26, 23, 24, 27, 22, 26, 24, 21...
```

```
## $ lseg12   <int> 9, 13, 8, 9, 10, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 9,...
## $ lseg13   <int> 8, 13, 9, 9, 10, 9, 9, 9, 9, 10, 10, 9, 9, 10, 10, 9, 9, 9...
## $ species  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

The data set is measurements on two species of biting fly which are very similar morphologically. Genus is Leptoconops. Species = 0 for L. torrens and species = 1 for L. carteri. Measurements are wing length and width (wingl, wingw), third palp length and width (thirdpl, thirdpw), fourth palp length (fourthpl) and length of antenna segments 12 and 13 (lseg12, lseg13). Note this is not leg length. It's the length of antenna segments. The goal here is to predict species from the measurements, if we can.

In this example, we want to fit a model with a binary predictor and a second one with a quantitative predictor. As there are no binary predictors, we will create one.

```
fly <- mutate(fly, thirdplbin = as.numeric(thirdpl > 37))
```

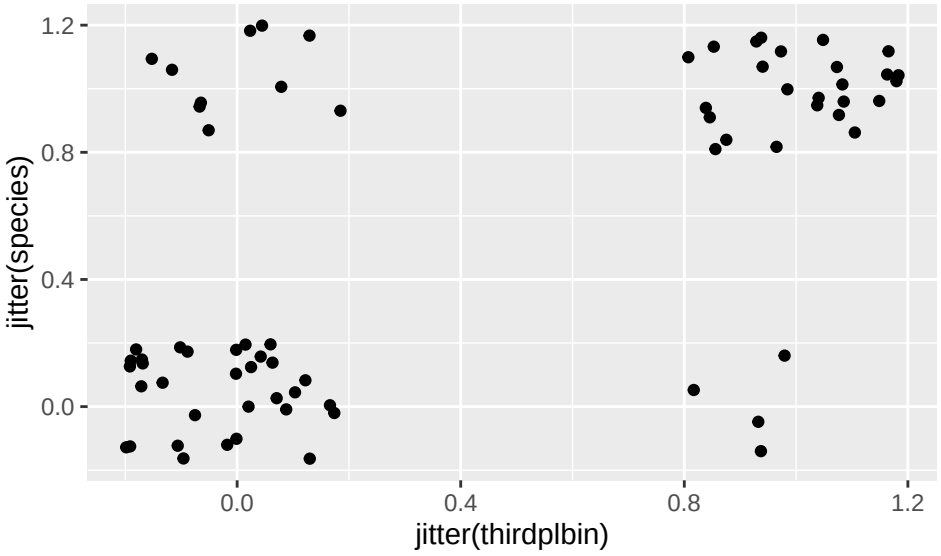
This creates a new variable that is 1 if thirdpl is greater than 37, and 0 if thirdpl is less than or equal to 37.

We want to examine the relationship between thirdplbin and species, which we can do with plots and a table.

Create an appropriate plot.

SOLUTION:

```
gf_point(jitter(species) ~ jitter(thirdplbin), data = fly)
```



Here is the table:

```
tally(~ species + thirdplbin, data = fly, format = "count")
```

```
##         thirdplbin
## species  0  1
##       0 31  4
##       1 10 25
```

Compute the probability of being species 1 (carteri) for flies with a thirdpl greater than 37. Compute the probability of being species 1 (carteri) for flies with a thirdpl less than or equal to 37.

SOLUTION:

```
25/29; 10/41
```

```
## [1] 0.862069
```

```
## [1] 0.243902
```

0.862 for thirdpl greater than 37, 0.244 for thirdpl less than or equal to 37.

> Compute the odds of being species 1 (carteri) for flies with a thirdpl greater than 37. Compute the odds of being species 1 (carteri) for flies with a thirdpl less than or equal to 37.

SOLUTION:

```
25/4;10/31
```

```
## [1] 6.25
```

```
## [1] 0.322581
```

The odds are 6.25 and 0.323 respectively.

> Compute the odds ratio of being species 1 for flies with a thirdpl greater than 37 relative to flies with a thirdpl less than or equal to 37.

SOLUTION:

```
(25/4)/(10/31)
```

```
## [1] 19.375
```

The odds ratio for P(carteri) is 19.4 for those with thirdpl > 37 compared to those flies that do not meet that criteria.

> WITHOUT fitting the model, use all your computations to report an estimated logistic regression line.

SOLUTION: Well, log(OR)= slope, and log(odds for < 37 group) = intercept. Those two values are:

```
log(19.4); log(10/31)
```

```
## [1] 2.96527
```

```
## [1] -1.1314
```

So the model would be logit(P(species=carteri))= -1.13 + 2.97(thirdplbin)

> Now fit the model. Report the fitted logistic regression line. Does it match what you predicted?

SOLUTION:

```
logm <- glm(species ~ thirdplbin, data = fly, family = binomial(logit))
msummary(logm)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.131      0.364   -3.11   0.0019 **
## thirdplbin     2.964      0.650    4.56  5.1e-06 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.041  on 69  degrees of freedom
## Residual deviance: 68.823  on 68  degrees of freedom
## AIC: 72.82
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is logit(P(species1)) = -1.131 + 2.964 (thirdplbin). It matches what we predicted.

> How do the three conditions (linearity, randomness, and independence) check out for your model?

SOLUTION: Our predictor is binary, so linearity is automatically satisfied. For randomness and independence, we don't really have information on that from the data set description. We need to be careful proceeding with inference. It turns out this WAS a random sample of flies from each species, so these conditions would all be satisfied.

Obtain a 95% confidence interval for the odds ratio based on the slope coefficient. Interpret the interval.

SOLUTION:

```
exp(confint(logm))
```

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

```
##                 2.5 %    97.5 %
## (Intercept) 0.150109  0.634895
## thirdplbin  5.921660 78.908101
```

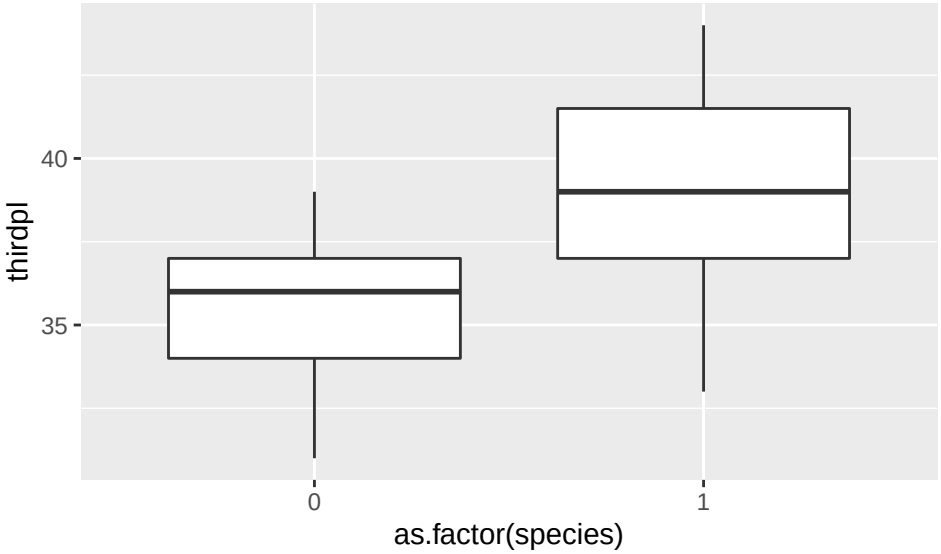
The 95% CI for the odds ratio is (5.92, 78.908). The odds for a fly with third pl greater than 37 to be L. carteri are between roughly 6 and 79 times the odds for a fly with third pl less than or equal to 37 to be L. carteri (500 to 7800 % increase in odds).

Now that we've fit a model with a binary predictor, we want to look at a quantitative one. So, let's explore the relationship between species and thirdpl.

```
favstats(thirdpl ~ as.factor(species), data = fly)
```

```
##   as.factor(species) min Q1 median   Q3 max    mean      sd  n missing
## 1                  0  31 34     36 37.0  39 35.3714 2.19740 35       0
## 2                  1  33 37     39 41.5  44 39.3143 2.83644 35       0
```

```
gf_boxplot(thirdpl ~ as.factor(species), data = fly)
```



Does there appear to be a relationship based on the descriptive statistics and boxplots generated?

SOLUTION: L. torrens appears to have lower values of thirdpl overall, so yes, there is some relationship.

Fit a logistic regression model predicting species using thirdpl. Report the fitted model.

SOLUTION:

```
logm2 <- glm(species ~ thirdpl, data = fly, family = binomial(logit))
msummary(logm2)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -26.030      6.594   -3.95  7.9e-05 ***
## thirdpl        0.699      0.178    3.94  8.2e-05 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.041  on 69  degrees of freedom
## Residual deviance: 62.436  on 68  degrees of freedom
## AIC: 66.44
##
## Number of Fisher Scoring iterations: 5
```

The fitted model is logit(P(Carteri)) = -26.030 + 0.699(thirdpl).

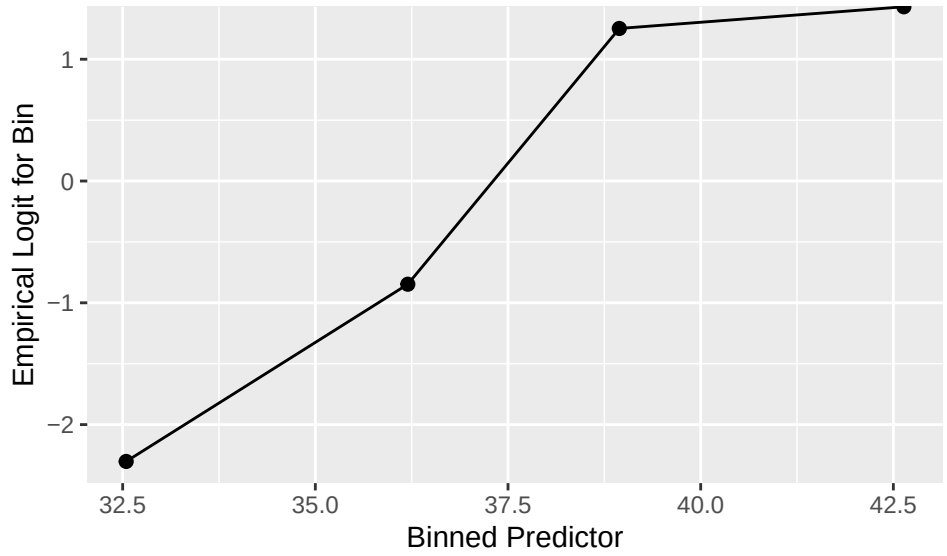
Interpret both coefficients from your fitted model.

SOLUTION: The intercept is the log odds of being L. carteri for flies with a thirdpl of 0. This is NOT actually meaningful here. The slope is the log of the odds ratio. For a one unit increase in thirdpl, the log odds of being L. carteri increases by 0.699. exp(0.699)=2.01. So, for a 1 unit increase in thirdpl, we expect the odds of being L. carteri to increase by basically 100%.

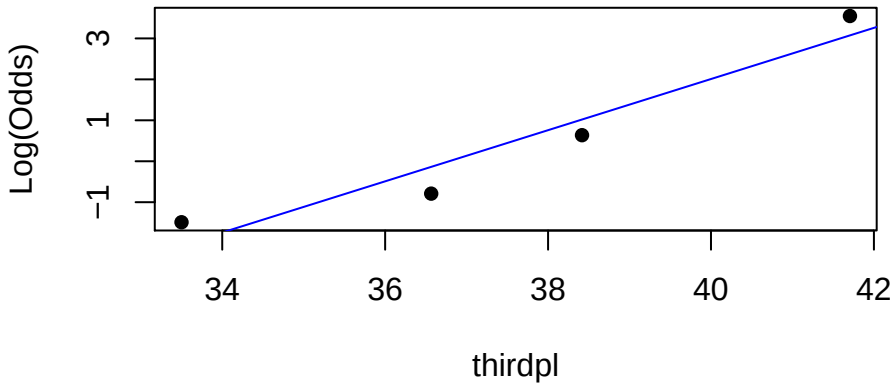
Use an appropriate plot to check the linearity condition for your model. Comment on whether the condition appears to hold.

SOLUTION:

```
with(fly, emplogitplot(species, thirdpl, 4))
```




```
emplogitplot1(species ~ thirdpl, ngroups = 4, data = fly)
```



It looks decently linear (not as nice as some of the class examples).

> Obtain a 95% confidence interval for the odds ratio based on the slope coefficient. Interpret the interval.

SOLUTION:

```
exp(confint(logm2))
```

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

```
##                    2.5 %      97.5 %
## (Intercept) 1.29575e-18 3.23630e-07
## thirdpl     1.49367e+00 3.02868e+00
```

The 95% CI for the odds ratio is (1.49, 3.03). We are 95% confident that the increase in odds for being L carteri is between a roughly 50% and 200% increase for each 1 unit increase in thirdpl.