DATA606 - Practice Question 8.15/8.17: Possums

Michael Y.

May 1, 2019

Possums!

The textbook has two problems using a dataset on Australian possums: 8.15 and 8.17

Photo credits: Greg Schechter, Arnaud Delberghe (via Flickr)

- Another classmate will be discussing 8.15 next week, so I'll present 8.17 today
- However, I have to touch on the description of the data, which is given in 8.15
- There is an issue with the data as stored in the datset, which makes it inconsistent with the textbook
- However, this can be fixed

The possum data set has 104 observations:

```
require(openintro)
data(possum)
head(possum)
```

```
##   site pop sex age headL skullW totalL tailL
## 1    1 Vic   m   8  94.1   60.4   89.0  36.0
## 2    1 Vic   f   6  92.5   57.6   91.5  36.5
## 3    1 Vic   f   6  94.0   60.0   95.5  39.0
## 4    1 Vic   f   6  93.2   57.1   92.0  38.0
## 5    1 Vic   f   2  91.5   56.3   85.5  36.0
## 6    1 Vic   f   1  93.1   54.8   90.5  35.5
```

```
tail(possum)
```

```
##     site   pop sex age headL skullW totalL tailL
## 99     7 other   f   3  93.3   56.2   86.5  38.5
## 100    7 other   m   1  89.5   56.0   81.5  36.5
## 101    7 other   m   1  88.6   54.7   82.5  39.0
## 102    7 other   f   6  92.4   55.0   89.0  38.0
## 103    7 other   m   4  91.5   55.2   82.5  36.5
## 104    7 other   f   3  93.6   59.9   89.0  40.0
```

Data Summary

```
summary(possum)
```

```
##       site          pop     sex         age            headL            skullW          totalL
##  Min.   :1.000   Vic  :46   f:43   Min.   :1.000   Min.   : 82.50   Min.   :50.00   Min.   :75.00
##  1st Qu.:1.000   other:58   m:61   1st Qu.:2.250   1st Qu.: 90.67   1st Qu.:54.98   1st Qu.:84.00
##  Median :3.000                     Median :3.000   Median : 92.80   Median :56.35   Median :88.00
##  Mean   :3.625                     Mean   :3.833   Mean   : 92.60   Mean   :56.88   Mean   :87.09
##  3rd Qu.:6.000                     3rd Qu.:5.000   3rd Qu.: 94.72   3rd Qu.:58.10   3rd Qu.:90.00
##  Max.   :7.000                     Max.   :9.000   Max.   :103.10   Max.   :68.60   Max.   :96.50
##                                    NA's   :2
##      tailL
##  Min.   :32.00
##  1st Qu.:35.88
##  Median :37.00
##  Mean   :37.01
##  3rd Qu.:38.00
##  Max.   :43.00
##
```

- The goal is to perform a logistic regression to predict whether a possum is from Victoria (Southeastern Australia) or from elsewhere in Australia based upon its **sex**, **head length**, **skull width**, **total length**, and **tail length**. (**Age** is not considered, perhaps because it contains 2 NA values, which would require imputation or exclusion.)

The textbook gives plots for 6 variables:

We can reproduce the plots (plus some color…)

- First three plots:

```
par(mfrow = c(1,3))
barplot(table(possum$sex),names.arg=c("Female","Male"),col=c("pink","blue"),xlab="sex_male",ylab="Frequency",main="Possums: Se
hist(possum$headL,breaks=22,col="lightblue",xlab="head length (in mm)",ylab="Frequency",main="Possums: Head Length")
hist(possum$skullW,breaks=22,col="lightgreen",xlab="skull width (in mm)",ylab="Frequency",main="Possums: Skull Width")
```

Reproduced Plots (2)

- Second row of plots:

```
par(mfrow = c(1,3))
hist(possum$totalL,breaks=22,col="brown",xlab="total length (in cm)",ylab="Frequency",main="Possums: Total Length")
hist(possum$tailL,breaks=22,col="red",xlab="tail length (in cm)",ylab="Frequency",main="Possums: Tail Length")
barplot(table(possum$pop),col=c("pink","blue"),xlab="population",ylab="Frequency",main="Possums: Geographic Distribution",widt
```

Logistic Regression (from 8.15)

```
p_reg1 <- glm(pop ~ sex + headL + skullW + totalL + tailL,data=possum,family=binomial)
summary(p_reg1)
```

```
##
## Call:
## glm(formula = pop ~ sex + headL + skullW + totalL + tailL, family = binomial,
##     data = possum)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8501  -0.3760   0.1182   0.5514   1.6430
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -39.2349    11.5368  -3.401 0.000672 ***
## sexm          1.2376     0.6662   1.858 0.063195 .
## headL         0.1601     0.1386   1.155 0.248002
## skullW        0.2012     0.1327   1.517 0.129380
## totalL       -0.6488     0.1531  -4.236 2.27e-05 ***
## tailL         1.8708     0.3741   5.001 5.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142.787  on 103  degrees of freedom
## Residual deviance:  72.155  on  98  degrees of freedom
## AIC: 84.155
##
## Number of Fisher Scoring iterations: 6
```

What's up with the signs on the coefficients?!

- There's a discrepancy here: Note that the **signs** on the coefficients on the above regression:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -39.2349    11.5368  -3.401 0.000672 ***
sexm          1.2376     0.6662   1.858 0.063195 .
headL         0.1601     0.1386   1.155 0.248002
skullW        0.2012     0.1327   1.517 0.129380
totalL       -0.6488     0.1531  -4.236 2.27e-05 ***
tailL         1.8708     0.3741   5.001 5.71e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- are **flipped** vs. the signs on the results shown in the textbook:

	Full Model			
	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	39.2349	11.5368	3.40	0.0007
sex_male	-1.2376	0.6662	-1.86	0.0632
head_length	-0.1601	0.1386	-1.16	0.2480
skull_width	-0.2012	0.1327	-1.52	0.1294
total_length	0.6488	0.1531	4.24	0.0000
tail_length	-1.8708	0.3741	-5.00	0.0000

- Why could this be?

Let's look closely at categorical variables:

```
barplot(table(possum$pop),width=c(1,1),space=2,col=c("pink","blue"),xlab="population",ylab="Frequency",main="Possums: Geograph
```

- Note that the order is **reversed**: In our data set, **"Vic"** appears on the **Left** while **"other"** appears on the right, but…
- In the subplot in the textbook, **"Not Victoria"** appears on the **left** (above a **"0"**), while **"Victoria"** is on the **right** (above a "**1**")

Assignment of Factors on categorical variables

- The textbook states that "**Victoria**" should be stored as "**1**", and "**Not Victoria**" should be stored as "**0**".

```
# Integer values of the factor
table(possum$pop)
```

```
##
##   Vic other
##    46    58
```

- However, the possum data table *actually* stores "**other**" (i.e., "**Not Victoria**") as a "**2**" rather than "**0**".

```
table(as.integer(possum$pop))
```

```
##
##  1  2
## 46 58
```

We can fix this:

- By creating a new population variable ("NEWpop") which is defined as (**2** minus the **original**),
- the value associated with **"Not Victoria"** becomes **"0"**, as expected per the textbook

```
possum$NEWpop = as.factor( 2 - as.integer(possum$pop))
table(possum$NEWpop)
```

```
##
##  0  1
## 58 46
```

Having forced "Not Victoria" to be "0" (rather than "2")

```
barplot(table(possum$NEWpop),
        col=c("blue","pink"),width=c(1,1),space=2,names.arg = c("Not Victoria","Victoria"),
        xlab="population",ylab="Frequency",main="Possums: Geographic Distribution")
```

- Observe that the values are now displayed in the expected sequence

We obtain the expected signs on the coefficients:

```
p_reg2 <- glm(NEWpop ~ sex + headL + skullW + totalL + tailL, data=possum, family=binomial)
summary(p_reg2)
```

```
##
## Call:
## glm(formula = NEWpop ~ sex + headL + skullW + totalL + tailL,
##     family = binomial, data = possum)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6430  -0.5514  -0.1182   0.3760   2.8501
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  39.2349    11.5368   3.401 0.000672 ***
## sexm         -1.2376     0.6662  -1.858 0.063195 .
## headL        -0.1601     0.1386  -1.155 0.248002
## skullW       -0.2012     0.1327  -1.517 0.129380
## totalL        0.6488     0.1531   4.236 2.27e-05 ***
## tailL        -1.8708     0.3741  -5.001 5.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142.787  on 103  degrees of freedom
## Residual deviance:  72.155  on  98  degrees of freedom
## AIC: 84.155
##
## Number of Fisher Scoring iterations: 6
```

Question 8.17 drops one variable:

- **Head Length** has been shown to be not significant, so it is dropped.
- The reduced model is as follows:

```
p_reg3 <- glm(NEWpop ~ sex +         skullW + totalL + tailL, data=possum, family=binomial)
summary(p_reg3)
```

```
##
## Call:
## glm(formula = NEWpop ~ sex + skullW + totalL + tailL, family = binomial,
##     data = possum)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8102  -0.5683  -0.1222   0.4153   2.7599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  33.5095     9.9053   3.383 0.000717 ***
## sexm         -1.4207     0.6457  -2.200 0.027790 *
## skullW       -0.2787     0.1226  -2.273 0.023053 *
## totalL        0.5687     0.1322   4.302 1.69e-05 ***
## tailL        -1.8057     0.3599  -5.016 5.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142.787  on 103  degrees of freedom
## Residual deviance:  73.516  on  99  degrees of freedom
## AIC: 83.516
```

8.17(a) - Write out the form of the model

$$logit(\hat{p_i}) = ln\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right) = \begin{array}{l} 33.5095 \quad -1.4207 \times sex_male_i - 0.2787 \times skull_width_i \\ \qquad\qquad +0.5687 \times total_length_i - 1.8057 \times tail_length_i \end{array}$$

- **Also identify which of the variables are positively associated when controlling for other variables.**
The only variable with a positive coefficient is `total_length`. This means that it is the only variable which is positively associated with a possum being from Victoria. (i.e., possoms from Victoria tend to be, on average, slightly **longer** than possums from elsewhere, but they have significantly **shorter tails**.)

- For the **numeric** variables, we can confirm this by segmenting the data by population:

```
by(data = possum[,c("skullW","totalL","tailL")], INDICES = possum$pop, FUN=colMeans)
```

```
## possum$pop: Vic
##   skullW   totalL    tailL
## 56.65435 87.46739 35.93478
## ------------------------------------------------------------------------
## possum$pop: other
##   skullW   totalL    tailL
## 57.06552 86.78793 37.86207
```

The average value for **"totalL"** is greater for the **"Vic"** segment, while the average values for **"skullW"** and for **"tailL"** are larger for the **"other"** segment.

Do we need to worry about the other categorical variable?

- Earlier we discovered that the way in which **"pop"** is stored in the dataset is not consistent with the description in the textbook, causing the signs on all the regression coefficients to flip.
- The other categorical variable is **"sex"** - according to the plot in the textbook, we assume sex=**male** corresponds to to "1", while sex=**female** corresponds to "0":

```
barplot(table(possum$sex),names.arg = c("Female","Male"),col=c("pink","blue"),xlab="sex_male",ylab="Frequency", main="Possums:
```

As it turns out, the factor levels behind "sex" are not what we expect

- We are expecting that **"Female"** should be stored as **"0"** while **"Male"** should be stored as **"1"**

```
# table of the "sex" factor:
table(possum$sex)
```

```
##
##  f  m
## 43 61
```

- ***However,*** the possum dataset *actually* stores **"Female"** as a **"1"** , while **"Male"** is stored as a **"2"** :

```
# integer values of the "sex" factor:
table(as.integer(possum$sex))
```

```
##
##  1  2
## 43 61
```

- Because the **order** is unchanged, the sign on the regression coefficient does not flip.

Are male possums less prevalent in Victoria?

```r
# On an ABSOLUTE basis:
table(possum$pop, possum$sex)
```

```
##
##          f  m
##   Vic   24 22
##   other 19 39
```

```r
# On a PERCENTAGE basis:
table(possum$pop, possum$sex)/length(possum$pop)
```

```
##
##                 f         m
##   Vic   0.2307692 0.2115385
##   other 0.1826923 0.3750000
```

- This explains the negative coefficient on **"sex"**: ***For this sample***, more Female possums were observed in Victoria, while more Male possums were observed elsewhere.
- However, is it reasonable that there should actually be *more than twice as many males as females* in the "other" region? Or, is it more plausible that this happens to be an unfortunate quirk of a small sample?

8-17(b) - Aussie possum in America

- "Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australis. However, the sign does indicate that the possum is:"
- **Male**; its **skull** is about 63mm wide; its **tail** is 37cm long, and its **total length** is 83cm.
- ***What is the reduced model's probability that this possum is from Victoria?***

```
this_possum = c(1, 1, 63, 83, 37)  # c(intercept, male, skullW, totalL, tailL)
rbind(coefs=p_reg3$coefficients,this_possum,product=p_reg3$coefficients*this_possum)
```

```
##             (Intercept)      sexm      skullW     totalL      tailL
## coefs          33.50947 -1.420672  -0.2787295  0.5687256  -1.805655
## this_possum     1.00000  1.000000  63.0000000 83.0000000  37.000000
## product        33.50947 -1.420672 -17.5599606 47.2042215 -66.809223
```

```
logit_p = sum(p_reg3$coefficients*this_possum)
logit_p
```

```
## [1] -5.076167
```

Solve for the probability

- The logit value is the dot-product of the coefficients and the variables
- For this example, the logit value is -5.0761666

$$logit(\hat{p_i}) = ln\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right) = -5.0761666$$

$$e^{logit(\hat{p_i})} = \frac{\hat{p_i}}{1-\hat{p_i}} = e^{-5.0761666} = 0.0062438$$

$$\hat{p_i} = \frac{e^{logit(\hat{p_i})}}{1+e^{logit(\hat{p_i})}} = \frac{0.0062438}{1+0.0062438} = \frac{0.0062438}{1.0062438} = 0.0062051$$

- In R, we can do the above calculations manually, or we can use the ***"logistic"*** function

```
phat <-logistic(logit_p)
phat
```

```
## [1] 0.006205055
```

How confident are you in the model's accuracy of this probability calculation?

- Given that the logistic value is quite low, 0.0062051, this model gives strong evidence that the possum in question is ***not*** from Victoria – *assuming that the model is correct*.
- The data that has been collected covers a relatively modest number of samples (n=104). The key question in determining correctness of the model is whether this sample is representative of the population of possums across the subject regions (Victoria, New South Wales, and Queensland, Australia.)
- One would expect that ***the wide discrepancy in the number of males and females by region*** as observed in ***this*** sample may suggest that ***this sample may not be representative of the actual population*** across each region.
- The result is that the measurements of the large number of male possums from the "other" (non-Victoria) regions may have biased these results in their favor.
- The model could be improved if the sampling were revised to obtain additional samples in the regions so as to more evenly balance out the number of males and females.