Basic Course on R: Logistic Regression Practical

Elizabeth Ribble*

18-24 May 2017

Contents

*emcclel3@msudenver.edu

1 Baby Data

1. Read in the data "R_data_January2015.csv" with a header and row names from the first column. Assign it to the object `babydata` and allow strings be converted to factors. Attach the data to the environment.

2. We would like to know if `smoking` predicts `Status`. Since `Status` is a binary variable (intellectual disability or normal brain development) we need to use logistic regression. Answer the following questions:

   (a) Write down the model with $\text{logit}(p)$, $\ln(p/(1-p))$, on the lefthand side, but instead of writing p write $P(\text{intellectual disability})$. Then write down the formula for this probability (the probabilty of having a baby with an intellectual disability).

   (b) If we run the model on the data as it is now, R will consider "normal brain development" as the event because it is second in the levels of `Status`:

```
levels(Status)

## [1] "intellectual disability"  "normal brain development"
```

   So we need to first change these factor levels so we treat "intellectual disability" as the event. Run the following trick:

```
table(Status)

## Status
##  intellectual disability normal brain development
##                       82                      108

newstatus <- factor(3-as.numeric(Status),
                    labels=c("normal brain development",
                             "intellectual disability"))
```

And to show that it worked:

```
levels(newstatus)

## [1] "normal brain development" "intellectual disability"

table(newstatus)

## newstatus
## normal brain development  intellectual disability
##                      108                       82
```

(c) Now run the regression model you set up above in R using `newstatus`. Then write down the model and probability of event with the estimates.

(d) Can `smoking` significantly predict `newstatus`? [Hint: use `summary`.]

(e) What is the probability of having a baby with an intellectual disability given the mother smokes?

(f) Our estimate of b_1 is the element in the 2nd row and 1st column of the coefficients from the `summary` call. What is the value of e^{b_1}? [Hint: use `exp`.]

(g) Is the e^{b_1} that you just calculated an odds ratio? How do you interpret it?

(h) What do you think e^{b_1} would have been if we didn't change the levels of `Status`? Re-run the model using `Status` to check your answer. How does it relate to your answer from (f)?

(i) There is another way to calculate an odds ratio without using logistic regression. Suppose we have the following 2×2 contingency table:

	event	no event
predictor yes	a	b
predictor no	c	d

then the odds ratio is $(a * d)/(b * c)$. Create a contingency table of `smoking` and `newstatus` [Hint: use `table`] and then calculate the odds ratio from that. Do you get the same answer as in (f)?

3. We would like to know if `smoking` and `vitaminB12` can be used to predict `newstatus`. Answer the following questions:

   (a) Set up (i.e. write down) the logit model and run it in R. Write the model with the estimates.

   (b) Can either variable significantly predict `newstatus`?

   (c) What is the probability of having a baby with an intellectual disability given the mother smokes and has a vitaminB12 level of 400? What is the probability of having a baby with an intellectual disability given the mother smokes and has a vitaminB12 level of 650?