Basic Course on R: Logistic Regression Practical Answers

Elizabeth Ribble*

17-21 December 2018

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.

```
babydata <- read.csv("R_data_January2015.csv",header=T,row.names=1)
attach(babydata)
```

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 probability of having a baby with an intellectual disability).
   The model is

$$\ln\left(\frac{P(\text{intellectual disability})}{1-P(\text{intellectual disability})}\right) = b_0 + b_1 * \text{smoking}$$

   and the probability of an event is

$$P(\text{intellectual disability}) = \frac{e^{b_0+b_1*\text{smoking}}}{1+e^{b_0+b_1*\text{smoking}}}.$$

   (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 and "normal brain development" as the baseline. Run the following to change all the 1s to 2s and 2s to 1s

```
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.

```
lrs1 <- glm(newstatus ~ smoking, family = binomial(logit))
lrs1$coef

## (Intercept)  smokingyes
##   -0.557015    1.367945
```

The model with estimates is

$$\ln\left(\frac{P(\text{intellectual disability})}{1-P(\text{intellectual disability})}\right) = -0.557 + 1.368 * \text{smoking}$$

and the probability of an event is

$$P(\text{intellectual disability}) = \frac{e^{-0.557+1.368*\text{smoking}}}{1+e^{-0.557+1.368*\text{smoking}}}.$$

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

```
summary(lrs1)$coef

##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.557015  0.1691109 -3.293786 0.0009884766
## smokingyes   1.367945  0.3859647  3.544224 0.0003937712
```

The p-value is less than 0.05 so `smoking` can be used as a predictor for `newstatus`.

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

```
mynew1 <- data.frame(smoking = factor("yes"))
logit_p1 <- predict(lrs1, newdata = mynew1)
p1 <- exp(logit_p1)/(1 + exp(logit_p1))
p1
```

```
##         1
## 0.6923077
```

(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`.]
We can exponentiate b_1 to get this value:

```
exp(summary(lrs1)$coef[2, 1])

## [1] 3.927273

## or
exp(1.3679)

## [1] 3.927095
```

(g) Is the e^{b_1} that you just calculated an odds ratio? How do you interpret it?
Since `smoking` is binary, $e^{b_1} = e^{1.37} = 4$ is the odds ratio. Thus we can say that smoking mothers have four times higher odds of having a child with an intellectual disability compared to non-smokers.

(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)?

```
lrs2 <- glm(Status ~ smoking, family = binomial(logit))
exp(summary(lrs2)$coef[2,1])

## [1] 0.2546296
```

Switching the event to the non-event inverses the odds ratio:

```
1/exp(summary(lrs1)$coef[2,1])

## [1] 0.2546296
```

So we could have just not changed the levels of `Status` and concluded that non-smoking women have 0.25 times smaller odds of having a baby with an intellectual disability compared to smoking women. But wasn't this more fun?

(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)?

```
tss <- table(smoking, newstatus)
tss[1,1]*tss[2,2]/(tss[1,2]*tss[2,1])

## [1] 3.927273
```

Same answer!

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.
The model is

$$\ln\left(\frac{P(\text{intellectual disability})}{1-P(\text{intellectual disability})}\right) = b_0 + b_1 * \text{smoking} + b_2 * \text{vitaminB12}$$

```
lrs3 <- glm(newstatus ~ smoking + vitaminB12,
            family = binomial(logit))
lrs3$coef

##  (Intercept)   smokingyes   vitaminB12
## -1.397551678  1.435938106  0.002087774
```

The model with estimates is

$$\ln\left(\frac{P(\text{intellectual disability})}{1-P(\text{intellectual disability})}\right) = -1.4 + 1.44 * \text{smoking} + 0.002 * \text{vitaminB12}$$

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

```
summary(lrs3)$coef

##                 Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.397551678 0.55509528 -2.517679 0.0118130962
## smokingyes   1.435938106 0.39134210  3.669266 0.0002432482
## vitaminB12   0.002087774 0.00130122  1.604474 0.1086096691
```

Yes, `smoking` is still significant (p-value < 0.05), but `vitaminB12` is not.

(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?

```
mynew <- data.frame(smoking = factor(c("yes", "yes")),
                    vitaminB12 = c(400, 650))
logit_p <- predict(lrs3, newdata = mynew)
p <- exp(logit_p)/(1+exp(logit_p))
p
```

```
##         1         2
## 0.7054726 0.8014592
```