Applied Logistic Regression

Juho Ruohonen

Spring 2017

1a: Cross-tabulating a Binary Response Variable and a Binary Predictor

In this exercise we have a case-control dataset that was collected to study the association between smoking and lung cancer. Here are the counts presented in a 2x2 contingency table:

```
(crosstab<-matrix(ncol = 2,c(647,2,622,27),dimnames =
  list(c("smoker","non-smoker"),c("cases","controls"))))
```

```
##            cases controls
## smoker       647      622
## non-smoker     2       27
```

In order to make manual calculation of the odds ratio easier, let's print a version of this table with the marginal totals shown:

```
addmargins(crosstab)
```

```
##            cases controls  Sum
## smoker       647      622 1269
## non-smoker     2       27   29
## Sum          649      649 1298
```

The odds of being a smoker among the cases is:

```
(odds.smoke.cases<-(647/649)/(1-(647/649)))
```

```
## [1] 323.5
```

The odds of being a smoker among the controls is:

```
(odds.smoke.controls<-(622/649)/(1-(622/649)))
```

```
## [1] 23.03704
```

The odds ratio of smoking among the cases vs. the controls is:

```
(OR.smoking<-odds.smoke.cases/odds.smoke.controls)
```

```
## [1] 14.0426
```

It's hard for me to say intuitively whether this difference looks statistically significant. But that is exactly what Pearson's chi-square test is used to measure.

1b: Pearson's Chi-Squared Test of Independence

As per the instructions, we will run the chi-square test on our contingency table **without** Yates' continuity correction:

```
chisq.test(crosstab, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  crosstab
## X-squared = 22.044, df = 1, p-value = 2.664e-06
```

The association is statistically significant at the highest (.001) level, as evidenced by the extremely low p-value.

2: More Chi-Square Testing

Same procedure, different data:

```
addmargins(crosstab2<-matrix(ncol=2, c(57,87,118,185),
  dimnames = list(c("exposed","non-exposed"),c("ill","healthy"))))
```

```
##             ill healthy Sum
## exposed      57     118 175
## non-exposed  87     185 272
## Sum         144     303 447
```

```
chisq.test(crosstab2, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  crosstab2
## X-squared = 0.016753, df = 1, p-value = 0.897
```

The chi-square test statistic is less than .2 here – in a 2x2 contingency table a value of at least 3.84 is required for statistical significance. Likewise the p-value of almost .9 plainly shows that the association doesn't even approach statistical significance.

3: Controlling for a Potential Confounder

Let us now suppose that the dataset above is stratified by the two levels of a potential confounder that is binary. The following are our two strata:

```
(stratum1<-matrix(ncol=2, c(16,50,80,165), dimnames=
list(c("exposed","non-exposed"),"Stratum 1"=c("ill","healthy"))))
```

```
##              Stratum 1
##               ill healthy
##   exposed      16      80
##   non-exposed  50     165
```

```
(stratum2<-matrix(ncol=2,
  c(41,37,38,20),dimnames=
    list(c("exposed","non-exposed"),"Stratum 2"=c("ill","healthy"))))
```

```
##              Stratum 2
##               ill healthy
##   exposed      41      38
##   non-exposed  37      20
```

The respective odds ratios of falling ill for the exposed vs. the non-exposed within each stratum are:

```
(OR.stratum1<-(16/80)/(50/165))
```

```
## [1] 0.66
```

```
(OR.stratum2<-(41/38)/(37/20))
```

```
## [1] 0.5832148
```

It immediately emerges that stratum is indeed having a confounding effect. Clearly, the apparent lack of effect of the exposure on illness was a false impression created by the uneven distribution of the strata. Another observation is that the "exposure" here appears to in fact be some preventative measure, since it reduces the rate of illness.

Let's now perform the chi-square test on each stratum individually, to find out if the effect of the exposure is statistically significant:

```
(x2.stratum1<-chisq.test(stratum1, correct = FALSE))
```

```
##
##  Pearson's Chi-squared test
##
## data:  stratum1
## X-squared = 1.7235, df = 1, p-value = 0.1892
```

```
(x2.stratum2<-chisq.test(stratum2, correct = FALSE))
```

```
##
##  Pearson's Chi-squared test
##
## data:  stratum2
## X-squared = 2.2925, df = 1, p-value = 0.13
```

The effect of the exposure is greater in Stratum 2, but it falls short of the statistical significance threshold in both strata. Chi-square value is 1.72 (p=0.19) for Stratum 1 and 2.3 (p=0.13) in Stratum 2, while the threshold of statistical significance is 3.84 for tables with these dimensions (df=1).

However, despite the lack of statistical significance, the exposure is clearly having at least some effect – a positive one. This fact was effectively obscured by the confounder earlier.

4: Common Odds Ratio

Let's now calculate the common odds ratio for the exposure across both strata. The common odds ratio of the strata is the *weighted average* of the stratum-specific odds ratios, in which strata with more observations are given more weight that ones with few observations. R has a function that does all this calculus with a single command, but for pedagogical reasons we will first try doing it by hand:

```
(commonOR.manual <- (((16*165)/311)+((41*20)/136)) / (((80*50)/311)+((38*37)/136)))
```

```
## [1] 0.6257834
```

This value seems intuitively believable given the partial odds ratios that we obtained in the previous exercise. Let's now have the corresponding R function calculate this, and hope we get the same result:

```
(juxtaposed.strata <- array(dim=c(2,2,2), c(stratum1,stratum2)))
```

```
## , , 1
##
##      [,1] [,2]
## [1,]   16   80
```

```
## [2,]   50  165
##
## , , 2
##
##      [,1] [,2]
## [1,]   41   38
## [2,]   37   20
```

```
mantelhaen.test(juxtaposed.strata, correct = FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  juxtaposed.strata
## Mantel-Haenszel X-squared = 3.9075, df = 1, p-value = 0.04807
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3925309 0.9976409
## sample estimates:
## common odds ratio
##         0.6257834
```

The result is indeed identical, so let us rejoice! I must say, though, that it took me a long time to figure out how to correctly apply the manual formula. The key thing that I eventually discovered was that the innermost parentheses are critically important for the operation to work correctly. In other words, $a_k d_k / n_k$ is wrong and $(a_k d_k)/n_k$ is right. This contradicts what I was taught about basic calculus in elementary school – I was told that parentheses are not necessary in division and multiplication and that the result will be the same regardless of the order in which you do the multiplications/divisions. It appears that I was shamelessly misled, causing me much wailing and gnashing of teeth.

5: Cochran-Mantel-Haenzel Test for Independence of Stratum and Outcome

The CMH test measures the probability that an observed difference in how different groups (strata) respond to a treatment is due to chance. The manual formula for computing the test statistic looks so cryptic and intimidating to a non-mathematician like me that I won't even try calculating it by hand. We'll just look at the Mantel-Haenszel chi-square value and the corresponding p-value reported by **mantelhaen.test()** in the previous section. Chi-square is 3.9 and p is just below .05. This implies that the difference between the strata is indeed significant, i.e., stratum is a significant confounder.[1]

As for an overall conclusion based on the CMH test result, mine doesn't differ significantly from the interpretation of the separate analyses of the strata. The preventive measure seems to be helping both strata to avoid the illness, although Stratum 2 benefits statistically significantly more. Since this is an illness that we're talking about, the practical implication for me is that if the preventive measure is not expensive, time-consuming or highly unpleasant, it should be recommended to everyone in the relevant population, regardless of their stratum.

6: Plotting Probabilities in Logistic Regression

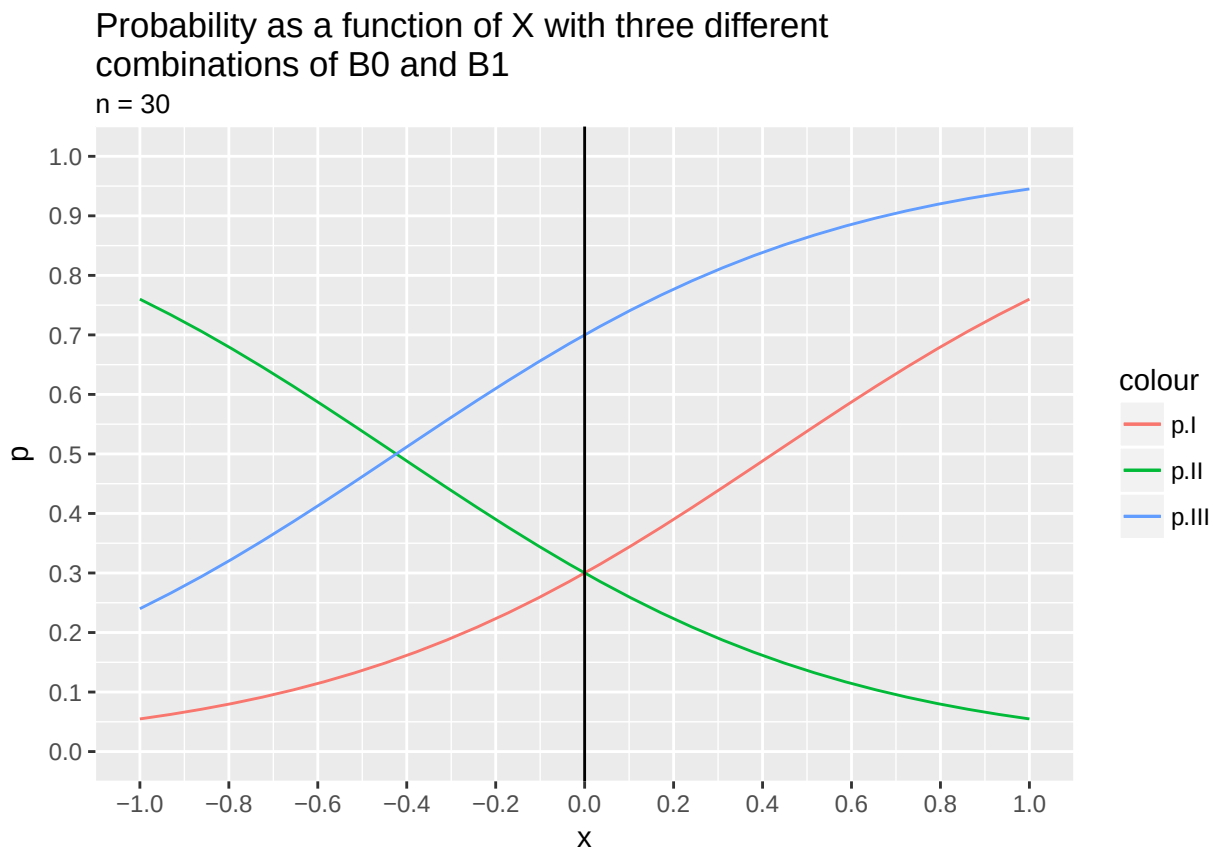
Logistic regression models the probability of the binary response being "success", but not directly – the natural logarithm of the odds corresponding to the probability, i.e., its *log-odds*, is used as a link between the probability and the predictors. The link function is a way around the inconvenient fact that probabilities have a restricted range between 0 and 1. An e-based logarithm, on the other hand, ranges between negative and

[1] If we run the CMH test *without* Yates' continuity correction, we get a p-value slightly above .06, which in principle is not statistically significant. The confidence interval remains entirely below 1, however.

positive infinity. This transformation of a probability into its log-odds is called the *logit transformation*. In solving the restricted range problem, the logit transformation makes the interpretation of logistic regression coefficients similar to the interpretation of the coefficients in linear regression – the sign of the coefficient immediately tells us whether whether a given predictor's effect on the outcome being a "success" is positive or negative.

Let us now plot the probability of "success" in three logistic regression models, all of which share $n = 30$ and a single continuous predictor $x = (i - 15.5)/14.5$, but each model having a different combination of constant term and predictor coefficient:

```
#The common parameters:
n <- 30
i <- 1:n
x <- (i-15.5)/14.5
#Model I
B0.I <- log(0.3/(1-0.3))
B1.I <- 2
logodds.I <- B0.I + B1.I*x[i]
p.I <- exp(logodds.I)/(1+exp(logodds.I))
#Model II:
B0.II <- log(0.3/(1-0.3))
B1.II <- -2
logodds.II <- B0.II + B1.II*x[i]
p.II <- exp(logodds.II)/(1+exp(logodds.II))
#Model III:
B0.III <- log(0.7/(1-0.7))
B1.III <- 2
logodds.III <- B0.III + B1.III*x[i]
p.III <- exp(logodds.III)/(1+exp(logodds.III))
#Storing all the data in one data frame (for plotting purposes):
OurData <- data.frame(i, x,
                      B0.I, B1.I, logodds.I, p.I,
                      B0.II, B1.II, logodds.II, p.II,
                      B0.III, B1.III, logodds.III, p.III)
#Plotting all three models on one X-Y plane:
library(ggplot2)
ggplot(data=OurData, aes(x=x, y=p.I)) +
  ggtitle("Probability as a function of X with three different \ncombinations of B0 and B1",
          subtitle = "n = 30") +
  ylab("p") +
  scale_y_continuous(limits=c(0,1), breaks=seq(from=0,to=1,by=0.1)) +
  scale_x_continuous(breaks=seq(from=-1,to=1,by=0.2)) +
  geom_line(aes(y=p.I,x=x,color="p.I")) +
  geom_line(aes(y=p.II,x=x,color="p.II")) +
  geom_line(aes(y=p.III,x=x,color="p.III")) +
  geom_vline(aes(xintercept=0))
```



This illustrates how the constant determines the value at which the graph intercepts the y axis, and the sign of the predictor coefficient determines whether the graph is ascending or descending. It is worth repeating here that even though the constant term is either **logit(0.3)** or **logit(0.7)** in each model, converting the logodds values back into probabilities also converts these constant terms back into the their input values, i.e., **logit(0.3)** and **logit(0.7)** become the y-intercept probabilities 0.3 and 0.7, respectively.

7: From Log-Odds to Probability and Back

If the log-odds of becoming ill are -2, then the probability of becoming ill is:

```
exp(-2)/(1+exp(-2))
```

```
## [1] 0.1192029
```

Conversely, if the probability of becoming ill is 0.119203, the log-odds of becoming ill are:

```
p <- 0.119203
round(log(p/(1-p)), digits=5)
```

```
## [1] -2
```
