Logistic Regression Models
Joseph M. Hilbe

PARTIAL REVISION CHAPTER 2.1

In *Logistic Regression Models* (2009, Chapman & Hall/CRC) , I begin by introducing the concepts of odds and odds ratios by means of a 2x2 and then by a 2x*k* table. This discussion covers pages 18-38 (chapter 2.1, 2.2). I later use the 2x2 table approach to explain the comparison between odds / odds ratios and risk / risk ratios (Pages 109-121; Chapter 5.5.2). These concepts and discussion is core to an understanding of the logistic regression model. It is also core to understanding the relationship of the binomial-logit and the Poisson models, which is extremely important when understanding how odds ratios are to be interpreted.

The manner by which I demonstrate the calculation of odds and odds ratios for 2xk tables and in the discussion in chapter 5.5.2 should probably have been used in the initial discussion of 2x2 tables. What is said in 2.1 is by no means mistaken, but I believe that the concepts could have perhaps been more clearly described by employing the method I used in 2.2 and 5.5.

Let me explain. Given the 2x2 table, using the format displayed in Stata output, we have in paradigm form:

y = response variable, or dependent variable
x = predictor, or independent variable

		x = 0	x = 1
y	0	A	B
	1	C	D

Odds of y=1 given x=1 : D/B
Odds of y=1 given x=0 : C/A
Odds ratio : [D/B]/[C/A] = (AD)/(BC)

		anterior = 0	anterior = 1
death	0	2504	2005
	1	67	120

This format differs from the one on page 18, but is the same. Since the response is typically displayed on the vertical axis and x on the horizontal, I use that format here. Note that in describing 2x*k* tables this format is employed.

Recall that a person experiencing an anterior site infarct (heart attack) is x=1, a person experiencing an inferior site infarct is x=0. A patient who dies within 48 hours of admission is y=1; a patient who survives beyond 48 hours of admission is y=0.

The odds of a patient experiencing an anterior infarct dying is 120/2005 = .05985037, or 6.0%
The odds of a patient experiencing an inferior infarct dying is 67/2504 = . 02675719 or 2.7%

The odds ratio of a patient experiencing an anterior site infarct dying compared to a patient experiencing an inferior site infarct is:

$$\frac{120/2005}{67/2504} = \frac{120*2504}{2005*67} = \frac{300480}{134335} = 2.2367961$$

Of course, we could divide the percentages we calculated above and come with the identical conclusion.

```
. di .05985037/.02675719
2.2367958
```

Therefore, patients experiencing an anterior infarct have a some 2.24 greater odds of dying than do inferior site infarct patients.

We enter the same data into Stata's editor and list it as:

```
. l
     +---------------------------+
     | death   anterior   count |
     |---------------------------|
  1. |     1          1     120  |
  2. |     1          0      67  |
  3. |     0          1    2005  |
  4. |     0          0    2504  |
     +---------------------------+
```

Performing a logistic regression on the data gives us the same results.

```
. logistic death anterior [fw=count], nolog

Logistic regression                               Number of obs   =       4696
                                                  LR chi2(1)      =      28.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -771.92263                       Pseudo R2       =     0.0179
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    anterior |   2.236796   .3476527     5.18   0.000     1.649411     3.03336
------------------------------------------------------------------------------
```

Later we discuss the difference between odds ratios and risk ratios. As a prelude, the risk of *y* given x=1 is D/(B+D), and of *y* given x=0 is C/(A+C). The risk ratio is then

$$\text{Risk Ratio} = \frac{D/(B+D)}{C/(A+C)} = \frac{D(A+C)}{C(B+D)} = \frac{AD + CD}{BC + CD}$$

Compare this with the odds ratio

$$\text{Odds Ratio} = \frac{AD}{BC}$$

When we interpret risk ratios we do so by saying, for example, that a patient having an anterior infarct is 2.6 times more likely to die within 48 hours of admission than an inferior infarct patient.

The difference is in the usage of "likely", which is probability language, not odds language. When the numerator (AD) is less than 10% the value of the denominator (BC), it is usually acceptable to interpret an odds ratio as if it is a risk ratio, and use "likely" in place of "odds of". But care must be taken when doing so. I tend to prefer always using "odds" language with logistic models, unless the numerator is truly rare; e.g. <1%.

CALCULATING RISK AND RISK RATIO

```
Risk if anterior
. di 120/(120+2005)  =   .05647059

Risk if inferior
. di 67/(67+2504)    =   .0260599

Risk ratio of death for anterior infarct patients compared to inferior
. di ( 120/(120+2005) ) / (67/(67+2504))
2.1669535

. di .05647059 / .0260599
2.1669534
```

We use a Poisson model with robust variance estimator to calculate risk ratios. Using the identical data for the logistic model, we obtain the same risk ratio as calculated by hand above.

```
. poisson death anterior [fw=count], nolog vce(robust) irr

Poisson regression                                Number of obs   =       4696
                                                  Wald chi2(1)    =      26.69
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood =  -776.2572                 Pseudo R2       =     0.0171
------------------------------------------------------------------------------
             |               Robust
       death |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    anterior |   2.166953   .3243484     5.17   0.000     1.616003    2.905741
------------------------------------------------------------------------------
```

Note: the difference between the odds and risk ratios; but most statisticians agree that in such a case we may interpret the odds ratio displayed for the earlier logistic model as a risk ratio. Just know that there is a difference, and why we are able to do so. Note also that I use the robust variance estimator rather than the model-based standard errors with the Poisson model. Why this is the case is explained in full in the text.