December 22, 2011

Risk, Odds, and their Ratios

Joseph Hilbe, *Arizona State University*

Risk, Odds, and their Ratios

The difference between risk ratios and odds ratio are commonly misunderstood by researchers. The difference is important when interpreting a logistic regression model when the coefficients have been exponentiated. Such exponentiated logistic coefficients are called odds ratio. The exponentiated coefficients for Poisson and negative binomial models are called relative rate ratios. Rate ratios are numerically the same thing to risk ratios, but the term is used when the response or dependent variable is a count.

To understand the difference between a risk ratio and an odds ratio, let's consider a 2x2 table looking like

	DELIVERY (in thousands)		
	delivered (0)	abortion (1)	TOTAL
MH			
0: no MH problems	30	25	55
1: MH problems	55	30	85
TOTAL	85	55	140

	0	1
0	A	B
1	C	D
	A+C	B+D

This is based on a study I had to review some time back, but the count of women subjects in the study is purely made up. I don't recall the actual counts. The study sample data consists of women throughout the United States who have gotten pregnant. The concern of the study is to predict whether a woman will experience mental health problems subsequent to the birthing experience.

The outcome (or response, or dependent variable - what you are interested in) is, therefore, mental health (MH) problems. I put it on the vertical axis. The explanatory predictor, or independent variable, is delivery (0) or abortion (1). We do not consider natural miscarriage.

The table partitions the count of women into those having MH problems subsequent to their birthing experience. Separate cells exist for the two types of *delivery* and two levels of *MH.* A schema has been placed to the right of the table, which can be used to show the calculations involved. I will refer to delivery as any non-abortion delivery.

The risk of MH problems if having an abortion is D/(B+D)
The risk of MH problems if having a delivery is C/(A+C)

The risk ratio of having MH problems following an abortion is [D/(B+D)]/[C/(A+C)] compared to a delivery. For the real numbers then, we have

The risk of a woman having mental health problems if she has an abortion is 30/55 = .545.
The risk of a woman having mental health problems if she delivers is 55/85 = .647
The risk ratio of subsequent MH problems following an abortion is

(30/55)/(55/85) =
.545/.647 = .842 = 84%

The odds and odds ratio is the same as risk and risk ratio except that the denominator is 0, no MH problems instead of the total for each group.

The odds of 1 (MH problems if having an abortion) is D/B
The odds of 0 (MH problems if having a delivery) is C/A

The odds ratio of having MH problems following an abortion is (D/B)/(C/A) = (A*D)/(B*C) compared to a delivery.

The odds ratio, which is typically stated as the odds of a woman having MH problems following an abortion (compared to delivering the baby), is

(30/25)/(55/30) = .6545 = 65%

If you have more than two groups of explanatory variable, as in the study, we can set it up as the table below. Note that I have split *delivered* above into two groups - normal delivery and unintended delivery. The key to remember is that one group or level is called the reference, and the statistical conclusions are based on it. Note that Stata uses the lowest value as the reference level by default. SAS and SPSS use the highest value as the reference by default.You can change the reference of course. For the above 2x2 table, level 0 was the reference (it usually is), and it is the denominator in the primary ratio. For our 3-level predictor, let's suppose that we choose having a normal delivery as the reference.

	DELIVERY (in thousands)			
	normal (1)	unintended (2)	abortion (3)	TOTAL
0 no MH problems	10	20	25	55
1 MH problems	15	40	30	85
TOTAL	25	60	55	140

There are two sets of risk, as well as odds, ratios. One compares 3 with 1, and the second compares 2 with 1 on the horizontal axis. So,

The risk of developing MH problems is (30/55)/(15/25) = .9090 = .91 (91%) for a woman having an abortion compared to having a normal delivery.

The odds of developing MH problems is (30/25)/(15/10) = .800 = .8 (80%) for a woman having an abortion compared to having a normal delivery.

The risk of developing MH problems is (40/60)/(15/25) = 1.11 = (111%) for a woman having an unintended pregnancy compared to having a normal delivery That is, women are 11% **more**

likely to have subsequent MH problems if they have an unintended pregnancy rather than a normal delivery..

The odds of developing MH problems is (40/20)/(15/10) = 1.333 = (133.3%) for a woman having an unintended pregnancy compared to having a normal delivery. That is, women have a 33 and a third percent **greater odds** of having subsequent MH problems if they have an unintended pregnancy compared to having a normal delivery.

For the second example above -- odds of abortion vs normal -- we can also interpret the relationship as: the odds of developing MH problems is 1/.8=1.25 or 25% greater for woman having a normal delivery compared to one having an abortion. The calculations to show this are: (15/10)/(30/25) = 1/25

Also -- you can switch references as well and make any group the reference. Just be careful to relate the correct groups. It is also important to remember that odds is not risk. With a risk ratio we can talk about probabilities and likelihoods - but not with odds ratios. Some epidemiologists and researchers make this mistake; in fact many do. In the paragraph directly above I cannot say that women having a normal delivery are 25% more likely, or 25% more probable, to have MH than those having an abortion, only that they have 25% greater odds.

Remember, I just made up the counts, or incidence rates. So they may not make good sense. Let's create a GLM model to estimate the odds and risk ratios of the above 2x3 data.

Using Stata's data editor, I created the following table. I can display it by using the *list*, or just *l*, command. Note that every cell is accounted for.

```
. list

     +-----------------------------+
     | health   delivery      count |
     |-----------------------------|
  1. |      0          1      10000 |
  2. |      0          2      20000 |
  3. |      0          3      25000 |
  4. |      1          1      15000 |
  5. |      1          2      40000 |
     |-----------------------------|
  6. |      1          3      30000 |
     +-----------------------------+


. save delivery
```

I will first model the data to determine the odds ratios for experiencing a mental health problem following delivery-type. Recall that there are three levels of delivery, with the first level declared as the reference. We will therefore obtain odds ratios for experiencing MH problems for 1) women having an abortion compared to those having a normal delivery, and 2) for women delivering an unplanned conception. Placing an "*i.*"as a prefix to the categorical predictor,

delivery, tells the software to factor the variable. The first level is designated as the reference by default. *Count*, the number of women in each cell in thousands, is entered into the model as a frequency weight. The *nolog* option suppresses a display of the iteration log, and *nohead* suppresses a display of the header statistics; e,g, deviance, Pearson, dispersion, log-likelihood, and so forth. A **logistic regression** is used to model the data.

```
. glm health i.delivery [fw=count], fam(bin) nolog nohead eform

------------------------------------------------------------------------------
             |                 OIM
      health | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    delivery |
          2  |   1.333333   .0207275    18.51   0.000     1.293321    1.374584
          3  |          .8   .0123935   -14.40   0.000     .7760742    .8246634
             |
       _cons |        1.5   .0193649    31.41   0.000     1.462522    1.538439
------------------------------------------------------------------------------
```

The odds ratios displayed in the above table are identical to those we calculated by hand a bit earlier. But here we get standard errors and 95% confidence intervals for each non-reference level of *delivery*. An odds ratio is displayed for the intercept as well, which is in fact incorrect. The exponentiation of the intercept is not a ratio. The intercept itself is understood as the value of the linear predictor when each predictor value in the model is zero.

The risk ratio is obtained using a **Poisson regression** with robust, or sandwich, standard errors. A robust, or Huber-White sandwich, variance estimator adjusts the standard errors for any correlation in the data that may be in excess of Poisson distributional assumptions. It does not change the coefficients, which when exponentiated are termed incidence rate ratios. To reiterate, the term rate is used here in place of risk when we are modeling count data.

```
. glm health i.delivery [fw=count], fam(poi) nolog nohead eform vce(robust)

------------------------------------------------------------------------------
             |               Robust
      health |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    delivery |
          2  |   1.111111   .0065734    17.81   0.000     1.098302     1.12407
          3  |   .9090909   .0058788   -14.74   0.000     .8976413    .9206865
             |
       _cons |          .6   .0030984   -98.92   0.000     .5939579    .6061036
------------------------------------------------------------------------------
```

Again, as with the odds ratios the values for the incidence rate ratios are identical to what we determined above by hand.

I should mention that the standard errors of the logistic model odds ratios are obtained using the delta method. They are not directly derived from the model variance-covariance matrix. The logistic model we used is displayed below, but parameterized to display coefficients, not odds ratios. The Stata matrix list e(V) command is used to display the variance-covariance matrix. Model standard errors are obtained by taking the square root of each diagonal term in the matrix. I shall create separate dummy variables for the three levels of delivery, using them in the model.

```
. tab delivery, gen(del)

   delivery |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2       33.33       33.33
          2 |          2       33.33       66.67
          3 |          2       33.33      100.00
------------+-----------------------------------
      Total |          6      100.00

. glm health del2 del3 [fw=count], fam(bin) nolog nohead

------------------------------------------------------------------------------
             |                 OIM
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        del2 |   .2876821   .0155456    18.51   0.000     .2572132     .318151
        del3 |  -.2231436   .0154919   -14.40   0.000    -.2535072   -.1927799
       _cons |   .4054651   .0129099    31.41   0.000     .3801621    .4307681
------------------------------------------------------------------------------

. matrix list e(V)

symmetric e(V)[3,3]
                 health:     health:     health:
                   del2        del3       _cons
 health:del2   .00024167
 health:del3   .00016667      .00024
health:_cons  -.00016667  -.00016667   .00016667
```

We take the square root of the diagonal terms of the matrix.

```
. di sqrt(.00024167)
.01554574

. di sqrt(.00024)
.01549193

. di sqrt(.00016667)
.01291007
```

The three values above are identical to the standard errors displayed in the model output. However, we cannot do this for the standard errors of odds ratios. Instead we use a formula based on the delta method:

$$SE_{OR} = \exp(\beta)*se$$

each standard error of the odds ratio is given as:

```
. di exp(_b[del2])*_se[del2]
.02072751

. di exp(_b[del3])*_se[del3]
.01239355

. di exp(_b[_cons])*_se[_cons]
.01936492
```

We compare the terms with the model output table, finding them to be identical.

```
. glm health del2 del3 [fw=count], fam(bin) nolog nohead eform

------------------------------------------------------------------------------
             |                 OIM
      health | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        del2 |   1.333333   .0207275    18.51   0.000     1.293321    1.374584
        del3 |         .8   .0123935   -14.40   0.000     .7760742    .8246634
       _cons |        1.5   .0193649    31.41   0.000     1.462522    1.538439
------------------------------------------------------------------------------
```

The confidence intervals are easy to calculate for any specified level of significance. For the standard 95% confidence interval, which represents a significance level of α=.05, is given as

$$\hat{\beta} \pm 1.96*SE$$

Given the model statistics for the second level of delivery as

```
health |    Coef.   Std. Err.      z      P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------
del2 |   .2876821   .0155456    18.51   0.000     .2572132     .318151
```

We can use the above formula given the coefficient and SE to calculate the confidence interval.

CONFIDENCE INTERVALS FOR COEFFICIENT: DEL2

```
. di .2876821 - 1.96 *   .0155456
.25721272

. di .2876821 + 1.96 *   .0155456
.31815148
```

which matches the confidence values given in the table of parameter estimates and related statistics.

For the confidence intervals for the odds ratios, simply exponentiate the model values above.

CONFIDENCE INTERVALS FOR ODDS RATIOS: DEL2

```
. di exp(.2876821 - 1.96 *   .0155456)
1.2933202

. di exp(.2876821 + 1.96 *   .0155456)
1.3745845
```

which are identical to the values displayed in the table of odds ratios and associated statistics above. The confidence intervals for the remaining coefficients and odds ratios in the model are calculated using the same methods.

The same logic obtains with respect to Poisson regression and risk or rate ratios. The standard errors and confidence intervals are calculated for coefficients and relative rate ratios in the same manner. Remember, however, that the standard errors for the risk or rate ratios were determined by using a robust or sandwich estimator.

POISSON REGRESSION: MODEL SEs

```
. glm health del2 del3 [fw=count], fam(poi) nolog nohead


------------------------------------------------------------------------------
             |                 OIM
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        del2 |   .1053605   .0095743    11.00   0.000     .0865953    .1241257
        del3 |  -.0953102        .01    -9.53   0.000    -.1149098   -.0757105
       _cons |  -.5108256    .008165   -62.56   0.000    -.5268287   -.4948226
------------------------------------------------------------------------------
```

CONFIDENCE INTERVALS: del2

```
. di _b[del2] - 1.96*_se[del2]
.08659494

. di _b[del2] + 1.96*_se[del2]
.12412609
```

POISSON REGRESSION: ROBUST SEs

```
. glm health del2 del3 [fw=count], fam(poi) nolog nohead vce(robust)

------------------------------------------------------------------------------
             |               Robust
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        del2 |   .1053605   .0059161    17.81   0.000     .0937652    .1169559
        del3 |  -.0953102   .0064667   -14.74   0.000    -.1079847   -.0826356
       _cons |  -.5108256    .005164   -98.92   0.000    -.5209469   -.5007044
------------------------------------------------------------------------------
```

ROBUST CONFIDENCE INTERVALS: del2

```
. di _b[del2] - 1.96*_se[del2]
.09376496

. di _b[del2] + 1.96*_se[del2]
.11695607
```

POISSON REGRESSION: RELATIVE RATE RATIOS

MODEL POISSON REGRESSION

```
. qui glm health del2 del3 [fw=count], fam(poi) nolog nohead
```

STANDARD ERRORS; DELTA METHOD

```
. di exp(_b[del2]) * _se[del2]
.01063808

. di exp(_b[del3]) * _se[del3]
.00909091

. di exp(_b[_cons]) * _se[_cons]
.00489898
```

Compare the above with the standard errors displayed in the table of relative rate ratios and associated statistics. They are identical.

```
. glm health del2 del3 [fw=count], fam(poi) nolog nohead eform

------------------------------------------------------------------------------
             |                 OIM
      health |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        del2 |   1.111111   .0106381    11.00   0.000     1.090455    1.132158
        del3 |   .9090909   .0090909    -9.53   0.000     .8914465    .9270845
       _cons |         .6    .004899   -62.56   0.000     .5904746    .6096791
------------------------------------------------------------------------------
```

The same method is used for calculating relative rate ratios with robust standard errors.

POISSON REGRESSION: ROBUST SE

```
. qui glm health del2 del3 [fw=count], fam(poi) nolog nohead vce(robust)
```

CALCULATE ROBUST SEs FOR RELATIVE RATE RATIOS

```
. di exp(_b[del2]) * _se[del2]
.00657345

. di exp(_b[del3]) * _se[del3]
.00587884

. di exp(_b[_cons]) * _se[_cons]
.0030984
```

ROBUST POISSON REGRESSION WITH RELATIVE RATE RATIOS

```
. glm health del2 del3 [fw=count], fam(poi) nolog nohead vce(robust) eform

------------------------------------------------------------------------------
             |               Robust
      health |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        del2 |   1.111111   .0065734    17.81   0.000     1.098302     1.12407
        del3 |   .9090909   .0058788   -14.74   0.000     .8976413    .9206865
       _cons |         .6   .0030984   -98.92   0.000     .5939579    .6061036
------------------------------------------------------------------------------
```

The robust standard errors are the same. The confidence intervals are calculated in the same manner as we did for odds ratios