BRIEF OVERVIEW ON INTERPRETING COUNT MODEL RISK RATIOS

An Addendum to

Negative Binomial Regression

Cambridge University Press (2007)

Joseph M. Hilbe

This short monograph is intended to augment my text, *Negative Binomial Regression* (2007, Cambridge University Press). Specifically, I provide the reader with an overview of the technical meaning of risk and risk ratio, and how to interpret the estimated incidence rate ratios that are displayed in the model output of Poisson and negative binomial regression. In addition, I discuss how to construct and interpret example interaction terms that are used for count response models. These items are either not addressed in the text, or not discussed in the detail that I now believe necessary in order to better understand the modeling of counts.

RISK AND RISK RATIO

We shall begin simply with a 2X2 table of counts. In terms of a model, we seek to understand the relationship of death within 48-hours of the onset of symptoms of having a myocardial infarction. In particular, we seek to understand the risk of death upon having an anterior-site infarct compared to the infarct being primarily located at another site on the heart.

The variables used to maintain information about death and infarct site are

Response => death : 1=died within 48 hours of onset of symptoms; 0= still alive
Predictor => anterior : 1=anterior-site infarct; 0=other-site infarct

The data is collected over a number of hospitals during the year, and can be tabulated as:

```
. tab death anterior

     Death |      1=anterior;
 within 48 |      0=other
 hrs onset |     Other   Anterior |     Total
-----------+----------------------+----------
         0 |     2,504      2,005 |     4,509
         1 |        67        120 |       187
-----------+----------------------+----------
     Total |     2,571      2,125 |     4,696
```

To see perhaps more clearly the way the values are stored in the dataset, we tabulate death on anterior without labels.

```
. tab death anterior, nolab

    Death |      1=anterior;
 within 48 |      0=other
 hrs onset |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,504      2,005 |     4,509
         1 |        67        120 |       187
-----------+----------------------+----------
     Total |     2,571      2,125 |     4,696
```

The risk or probability of death for a patient having an anterior infarct is 120/2125 = .05647059. or approximately 5.6%.

The risk or probability of death for patient having an other-site infarct is 67/2571 = .0260599 or approximately 2.6%.

The ratio of the risk of death for anterior-site patients compared to the risk of death for other-site patients is .05647/.02606 = 2.1669225

This means that anterior-site infarct patients are over twice as likely to die within 48 hours of the onset of symptoms as are other-site infarct patients. The risk of anterior-site patients dying is 2.17, or 217%, greater than for other-site patients. Statisticians many times refer to this ratio as an **incidence rate ratio**. It is a ratio based on the rate or incidence of counts. It can also be thought of as a ratio of ratios; i.e. the base ratio is the incidence rate of counts having some characteristic or property out of a group consisting of the population of subjects or items from which the counts are a part. From among those patients having an anterior-site infarct, 120 died out of the study population of 2125 --- a ratio of 120:2125. Likewise, from among those patients having an other-site infarct, 67 died out of the study population of 2571 --- a ratio of 67:2571. To compare anterior with other-site infarct deaths, we take the ratio of the two ratios, for the incidence rate ratio (IRR).

The IRR can be calculated directly from the table by using the following Stata code:

```
. di (120/2125) / (67/2571)
2.1669535
```

Note that it is traditional that the ratios of 120/2125 and 67/2571 are simply referred to as risks. That is, 120/2125 is the risk of death for anterior-site infarct patients; 67/2571 is the risk of death for other-site infarct patients. The risk ratio is then the ratio of the two risks. However, remember that in fact the risks are themselves ratios.

Poisson regression is the basic count model used by statisticians to relate a count variable with one or more explanatory predictors. That is, the counts in the response variable are to be understood by virtue of one of more predictors. Here we have only one predictor, *anterior*, and a response of *death*. In this situation *death* is not really a count, but rather a binary variable: *died* or *not-died*. For understanding the nature of risks it makes no difference, but it is important when comparing basic risk ratios to odds ratios.

I shall use the Generalized Linear Models (GLM) method of implementing a Poisson regression using Stata. *nolog* indicates that no iteration log is displayed, *eform* indicates that the Poisson coefficients are exponentiated, and *nohead* means that no other statistics other then the risk ratio and its standard error, and z- and p-values and confidence intervals are displayed.

POISSON REGRESSION with exponentiated coefficient

```
. glm death anterior, nolog fam(poi) eform nohead
```

death	IRR	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
anterior	2.166953	.3304779	5.07	0.000	1.607069	2.921895

The IRR value is identical to what we calculated by hand from the table of counts.

COMPARE TO ODDS RATIO

Many times one wishes to calculate an odds ratio from a table like we used to construct a risk ratio, In fact, the two statistics – risk ratio and odds ratio – are comparable. Moreover, far too many times statisticians will interpret odds ratios as if they were risk ratios. For most cases this can be a serious error.

Let's look at the same table as we used for the risk ratio

```
tab death anterior, nolab
```

Death within 48 hrs onset	1=anterior; 0=other 0	1	Total
0	2,504	2,005	4,509
1	67	120	187
Total	2,571	2,125	4,696

The odds of an event occurring is simply the probability of an event occurring divided by the probability of an event not occurring. When an event is characterized as

success/failure, or died/not-died, statisticians record the relationship as 1/0. 1=success (or whatever one is interested in); 0=not a success. Therefore, an odds ratio can be schematized as

$$OR = \frac{p}{1-p}$$

where p is the probability of 1, and 1-p is the probability of 0; ie 1-1. For the table above, we have for anterior-site infarct patients:

ODDS OF DEATH FOR ANTERIOR-SITE INFARCT PATIENTS

120/2005 = .05985037 = 6.0%

or among anterior-site patients, there are 120 1's and 2005 0's.

For other-site patients, we have the odds:

ODDS OF DEATH FOR OTHER-SITE INFARCT PATIENTS

67/2504 = .02675719 = 2.7%

The odds ratio is the ratio of the two odds: odds of anterior-site over odds of other-site

ODDS RATIO OF DEATH FOR ANTERIOR-SITE VS OTHER SITE INFARCTS

.05985037 / .02675719 = 2.2367958

We model a binary logistic regression using the same command as we did for Poisson regression, except that we use the binomial family instead of the Poisson. The default link of the binomial family is the logit link, which produces what is typically called logistic regression.

```
. glm death anterior, nolog fam(bin) eform nohead

------------------------------------------------------------------------------
             |                 OIM
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    anterior |   2.236796   .3476532     5.18   0.000      1.64941    3.033361
------------------------------------------------------------------------------
```

Note that the values of the odds ratio and risk ratio are closely the same. This happens when the numerator of the ratio is small compared to the numerator. When the numerator

is less than 10% the denominator, the odds and risk ratios are close, as they are here. In this case it is regarded as permissible to interpret an odds ratio as a risk ratio.

What is the advantage of using risk ratios rather than odds ratios? For some the use of the terms, risk, or likely, is easier to understand than odds ratio. For instance, for the above logistic regression model, the correct interpretation of the odds ratio is that

> Anterior-site patients have a near 2 and a quarter greater odds of death (within 48 hours of the onset of symptoms) than other-site patients.

However, most researchers would prefer to say that

> Anterior-site patients have a near 2 and a quarter greater probability or likelihood of death (within 48 hours of the onset of symptoms) than other-site patients.

The latter is risk-language, not odds-language. In this situation, though, because of the low incidence rate of death for anterior-site patients, we may use risk-language. But most logistic models do not have a response with such a low incidence rate, and should be interpreted as odds ratios instead. Unfortunately we find a misuse of the terms too many times in the literature.

COUNT DATA

Thus far, I have addressed risk and risk ratios in terms of binary variables. I now turn to the modeling of a range of counts.

A count is generally understood as a list of non-negative integers, ie. 0, 1, 2, 3, …, n Usually when one is modeling counts, the number of counts is between 8 and 40. When there are a large number of different values of counts, it may be preferable to model as if it were continuous. For example, if the response variable has several hundred different values ranging from 0 to 1000, modeling as a lognormal model may be preferred to modeling as a count.

Another consideration arises when real numbers (with decimal places) are modeled as counts. The Poisson distribution upon which the Poisson regression model is based assumes that each value is an integer, but will allow modeling of real numbers. The model is still valid, but the distributional assumptions upon which the Poisson model is based have been violated. On the other hand, Poisson regression is robust to integer violations, and many statisticians do on fact use count models such as Poisson and negative binomial regression with non-integer data. What is not permissible however, is the modeling of negative data.

Suppose that we have data on the number of visits that patients make to the doctor each year. The number can range from 0 (no visits) to 20 or more. This is an ideal type of count variable. We also can have predictors which enable us to predict -- and understand – what may contribute to a patient visiting the doctor more often than others. Predictors can be *age*, meaning that perhaps older patients visit the doctor more often than younger ones, or female patients may visit more than males. A host of other possible predictors of doctor visits can be imagined. When they are part of an actual study, the count of visits can be modeled using a variety of count models.

TYPES OF PREDICTORS

For our purpose, there are three varieties of predictor:

Binary	1/0: male/female, yes/no, died/not-died, graduated/not-graduated
Categorical	ordered or unordered categories: 1, 2, 3, 4 …, J A,B,C,D,F freshman, sophomore, junior, senior Agree, Somewhat agree, Neutral, Somewhat disagree, Disagree
Continuous	height, weight, systolic blood pressure, age

Each type of predictor has a different type of association with the response term, which is reflected in the interpretation of the incidence risk or rate ratio (also known as relative risk ratio, or relative rate ratio – all meaning the same).

INTERPRETATION OF RISK RATIOS

I shall use the **rwm_1980x** dataset for an example of how to interpret risk ratios. The model I shall develop consists of predictors having all three types of variable as listed above. The data come from a German Health Data Survey which recorded subject visits from 1984-1988.

CHECKING MODEL VARIABLES

The number of visits to the doctor during a given year by an individual patient is (partially) tabulated

```
. tab docvis

     docvis |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      7,572       38.61       38.61
          1 |      2,582       13.17       51.78
          2 |      2,357       12.02       63.80
          3 |      1,858        9.48       73.28
          4 |      1,137        5.80       79.08
          5 |        792        4.04       83.11
          6 |        683        3.48       86.60
          7 |        393        2.00       88.60
          8 |        370        1.89       90.49
          9 |        196        1.00       91.49
         10 |        347        1.77       93.26
         11 |        130        0.66       93.92

         80 |          1        0.01       99.97
         82 |          1        0.01       99.97
         84 |          1        0.01       99.98
         90 |          2        0.01       99.99
        100 |          1        0.01       99.99
        121 |          1        0.01      100.00
------------+-----------------------------------
      Total |     19,609      100.00
```

We can see that nearly 40% of the patients never visited the doctor during any of the 5 years of the study. One patient, however, visited 121 times for one of the years of the study. We can identify that patient and obtain his or her patient ID and year using the following code.

```
. l id year age if docvis==121

      +-------------------+
      |   id   year   age |
      |-------------------|
5293. | 1885   1984    38 |
      +-------------------+
```

The patient with ID 1885 made 121 doctor visits during 1985.

To see how many visits that patient had in other years,

```
. l id year age docvis if id==1885

      +---------------------------+
      |   id   year   age   docvis |
      |---------------------------|
5293. | 1885   1984    38      121 |
      +---------------------------+
```

Patient 1885 was part of the survey for only one year, in 1984. For the patient having 100 visits during a given year, we have:

```
. l id year age if docvis==100

      +------------------+
      |   id   year   age |
      |------------------|
 5294. | 1886   1985    57 |
      +------------------+

. l id year age docvis if id==1886

      +---------------------------+
      |   id   year   age   docvis |
      |---------------------------|
 5294. | 1886   1985    57      100 |
 5295. | 1886   1986    58       22 |
 5296. | 1886   1987    59        0 |
 5297. | 1886   1988    60       37 |
      +---------------------------+
```

Note that the patient visited 100 times in 1985, but none in 1987. Note also that *age* increases in value by one year for every year a patient is in the study, as is expected. The values, however, are not independent within ID groupings. This may result in a problem if *age* is considered as an independent predictor of visits. Since we are not discussing the fit of models, I shall overlook this potential problem.

POISSON MODEL OF NUMBER OF VISITS TO DOCTOR

We model visits to the doctor by:

outwork: binary : 1=patient out of work
female : binary : 1=female
married : binary : 1=married
age : continuous (25-64)
edlevel : categorical 1=Not HS grad, 2=HS grad, 3=Coll/Univ, 4=Grad school

Create separate indicator (or dummy) variables for each category of *edlevel*. Each will be formatted as 1/0, with 1=membership in that category or level.

```
. tab edlevel, gen(edlevel)
```

With *edlevel1* as the reference, model using Poisson regression. At this point I shall display only the table of risk ratios and associated standard errors and confidence intervals. Summary and fit statistics are of foremost concern to assessing the fit and appropriateness of the model, but take us beyond the limited scope of this monograph. I shall acknowledge, however, that the Poisson model below is severely overdispersed, and is hence not a well fitted model.

```
. glm docvis outwork female married age edlevel2-edlevel4, nolog fam(poi)
nohead eform

------------------------------------------------------------------------------
             |                 OIM
      docvis |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |   1.248022   .0116171    23.80   0.000      1.22546    1.271001
      female |   1.211805   .0109699    21.22   0.000     1.190494    1.233497
     married |   .9077947    .008746   -10.04   0.000     .8908137    .9250994
         age |   1.019691   .0003841    51.77   0.000     1.018938    1.020444
    edlevel2 |   .9230602   .0164689    -4.49   0.000     .8913396    .9559098
    edlevel3 |   .8017075   .0131623   -13.46   0.000     .7763205    .8279246
    edlevel4 |   .6785868   .0140743   -18.69   0.000     .6515548    .7067404
------------------------------------------------------------------------------
```

INTERPRETATION

OUTWORK

The effect of a patient being out of work is to increase the expected number of visits to the doctor in a year by some 25%.

FEMALE

Females visit the doctor throughout the year some 21% more than males.

MARRIED

Married patients visit the doctor nearly 10% less often each year than do unmarried patients.

AGE

Each additional year of age is associated with an estimated 2% increase in doctor visits. A 10 year increase in the age of the patient is associated with an incidence rate ratio of 1.019691^10 = 1.2153066. This relates to a some 22% increase in the number of doctor visits. That is, patients 10 years older visit the doctor 22% more of the time.

EDLEVEL2

Patients whose highest level of education is high school graduation visit the doctor 8% less than patients who are high school drop outs.

EDLEVEL3

Patients whose highest level of education is attendance at a college or university visit the doctor some 20% less than patients who are high school drop outs.

EDLEVEL4

Patients whose have attended graduate school visit the doctor some 32% less than patients who are high school drop outs.

INTERACTIONS

I did not discuss interactions in the text. The reason is that I assumed that readers would have a background in this area of statistics. But in case someone is not familiar with interactions for non-linear models and how they are interpreted, I will give a brief overview of the construction and interpretation of interactions in count models.

The most common interactions are
- Binary X Binary
- Binary X Continuous
- Categorical X Continuous

Also used in research, but not as much as the three above listed types of interaction, are the Binary X Categorical, Categorical X Categorical, and Continuous X Continuous interactions. Since this is an overview, I shall here discuss only the three most commonly used types of interaction.

Suppose that we have a model consisting of a count response, a binary risk factor --- or predictor of primary interest --- and another predictor. When it is clear that the relationship of the response and risk factor is influenced by a third variable, we say that the third variable is a confounder of the relationship of response and risk factor. If there is a statistically significant relationship between the two predictors and a response, we say that there is an interaction between the two predictors. I shall use the same German Health Survey data as used earlier for constructing examples of each but the last interaction, the Categorical X Continuous interaction. In addition, for simplicity I shall exclude *edlevel* from the example models.

BINARY X BINARY INTERACRTIONS

A possible interaction can be identified if we exclude a predictor from a model and note that the risk ratio (or coefficient) of another predictor is greatly affected. For example, let us run the model of doctor visits on *outwork* and *female*. For the purpose of examining the nature of interactions I do not include other predictors, however, their inclusion would make no difference to our results.

```
. keep docvis outwork female

. glm docvis outwork female, nolog fam(poi) nohead

------------------------------------------------------------------------------
             |                 OIM
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |   .3534121   .0089587    39.45   0.000     .3358535    .3709708
      female |   .2016516    .008996    22.42   0.000     .1840198    .2192834
       _cons |   .9100899   .0063533   143.25   0.000     .8976377    .9225421
------------------------------------------------------------------------------
```

Drop *female* from the model and observe the change in parameter estimate in outwork, if any. If there is substantial change, this indicates a possible interactive effect.

```
. glm docvis outwork, nolog fam(poi) nohead

------------------------------------------------------------------------------
             |                 OIM
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |   .4426627   .0080554    54.95   0.000     .4268744    .4584509
       _cons |     .98046      .0054   181.57   0.000     .9698762    .9910437
------------------------------------------------------------------------------
```

There appears to be a sizeable alteration in values. We therefore suspect and test for an interactive effect. We do this by multiplying *outwork* and *female*, creating a new variable which we call *oxf*.

```
. gen byte oxf = outwork * female
```

oxf is included in the model, together with both main effects, i.e., with both *outwork* and *female*. You MUST have the main effects in the model, just as we keep lower terms in an ANOVA when there are interactions. The same logic obtains.

```
. glm docvis outwork female oxf, nolog fam(poi) nohead

------------------------------------------------------------------------------
             |                 OIM
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |   .6678495   .0139887    47.74   0.000     .6404322    .6952668
      female |   .3826608   .0109585    34.92   0.000     .3611827     .404139
         oxf |  -.4919359   .0176983   -27.80   0.000    -.5266239    -.457248
       _cons |   .8388561   .0070617   118.79   0.000     .8250153    .8526968
------------------------------------------------------------------------------
```

In terms of incidence rate ratios, we have

```
. glm docvis outwork female oxf, nolog fam(poi) nohead eform

------------------------------------------------------------------------------
             |                 OIM
      docvis |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |   1.950039   .0272785    47.74   0.000     1.897301    2.004244
      female |   1.466181   .0160671    34.92   0.000     1.435026    1.498012
         oxf |   .6114415   .0108215   -27.80   0.000     .5905955    .6330234
------------------------------------------------------------------------------
```

To determine the meaning of the interaction, we use the formula:

$$\beta_{\text{BxB}} = \beta_1 + \beta_3 x$$

And for incidence rate ratios:

$$IRR_{BxB} = \exp[\beta_1 + \beta_3 x]$$

For our interaction we have:

β_1 = coefficient on outwork
β_2 = coefficient on female
β_3= coefficient on interaction

OUT OF WORK FEMALES

$$IRR_{BxB} = \exp[0.6678497 - 0.4919363*female=1]$$

$$= 1.1923348$$

INTERPRETATION
Among females, those who are out of work visit the doctor 1.19 times more often than those who work.

and

OUT OF WORK MALES

$$IRR_{BxB} = \exp[0.6678497 - 0.4919363*female=0]$$

$$= 1.9500396$$

INTERPRETATION
Among males, those who are out of work are nearly twice as likely to visit the doctor than those who work.

and also, using the formula: $IRR_{BxB} = \exp[\beta_2 + \beta_3 x]$

WORKING MALES

$$IRR_{BxB} = \exp[0.382661 - 0.4919363*outwork=1]$$

$$= .89648358$$

INTERPRETATION
Among those who are out of work, females are 10% less likely to visit the doctor than males.

WORKING FEMALES

$$IRR_{BxB} = \exp[0.382661 - 0.4919363*outwork=0]$$

$$= 1.4661794$$

INTERPRETATION
Among those who are working, females are 47% more likely to visit the doctor than males.

Note that each combination of *outwork* and *female* with respect to *docvis* has an interpretation. There is no ONE interpretation of the interaction, and the coefficient, or corresponding IRR is not value used to directly interpret the interaction.

STANDARD ERROR AND CONFIDENCE INTERVAL

I shall detail the calculation of the standard errors and confidence intervals for the various interactions in the following section. The method used for Binary X Binary interactions is identical to that used for Binary X Continuous interactions. Likewise, other interactions use the same method.

BINARY X CONTINUOUS INTERACTIONS

Let us suppose we propose an interaction between married and age. We create such an interaction using the following code, with *mxa* as the interaction term.

```
. gen mxa = married * age
```

We model the data including the *oxf* interaction. However, if we discovered an apparent interaction between *outwork* and *married*, for example, we should not employ it as an interaction – unless we drop *oxf*. Having one term in two interactions confounds the relationship so much that it is very difficult to interpret. I recommend that it not be done. If you have more than one interaction in a model, be certain that no main effect term is used in both interactions. Again, it can be done, but the interpretation is tedious.

```
. glm docvis outwork female oxf married age mxa, nolog fam(poi) eform nohead

------------------------------------------------------------------------------
             |                 OIM
      docvis |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |    1.68131    .024236    36.04   0.000     1.634473    1.729489
      female |   1.470328   .0162411    34.90   0.000     1.438839    1.502507
         oxf |   .6347213   .0113969   -25.32   0.000     .6127722    .6574565
     married |   .8273458    .030434    -5.15   0.000     .7697957    .8891982
         age |   1.018261   .0006636    27.77   0.000     1.016961    1.019562
         mxa |   1.003087    .000785     3.94   0.000      1.00155    1.004627
------------------------------------------------------------------------------
```

We may still interpret the main effect terms in the same manner as we did for the Binary x Binary interaction, but the interaction term is interpreted differently. In fact, we do not use the *mxa* coefficient or IRR directly when interpreting the interactive relationship.

In order to calculate the Binary X Continuous interaction, we first must construct a table of parameter estimates, or coefficients.

```
. glm docvis outwork female oxf married age mxa, nolog fam(poi) nohead

------------------------------------------------------------------------------
             |                 OIM
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     outwork |   .5195731    .014415    36.04   0.000     .4913202    .5478259
      female |   .3854858   .0110459    34.90   0.000     .3638362    .4071354
         oxf |  -.4545693   .0179558   -25.32   0.000     -.489762   -.4193767
     married |  -.1895326   .0367851    -5.15   0.000    -.2616301   -.1174351
         age |   .0180964   .0006517    27.77   0.000     .0168191    .0193736
         mxa |   .0030823   .0007826     3.94   0.000     .0015485    .0046161
       _cons |   .1022505   .0301124     3.40   0.001     .0432313    .1612697
------------------------------------------------------------------------------
```

Next, use the following formula to calculate the IRR for *mxa*:

$$IRR_{BxC} = \exp[\beta_1 + \beta_3 x]$$

Where IRR_{BxC} is the interaction of Binary X Continuous predictors, β_1 is the coefficient of the binary term, β_3 is the coefficient of the interaction, and x is the value of the continuous predictor.

For our model we have:

$$IRR_{mxa} = \exp[\beta_1 + \beta_3 age]$$

$$IRR_{mxa} = \exp[-.1895326 + .0030823 * age]$$

Note that there is a different IRR value for each value of age. For ages 30, 40, 50, and 60 we have:

AGE = 30

```
. di exp(-.1895326 + .0030823*30)
.90749829
```

AGE = 40

```
. di exp(-.1895326 + .0030823*40)
.93590566
```

AGE = 50

```
. di exp(-.1895326 + .0030823*50)
.96520226
```

AGE = 60

```
. di exp(-.1895326 + .0030823*60)
.99541594
```

The next problem is in determining the standard errors (SEs) and confidence intervals (CIs) for the various instances of the interaction. The SEs and CIs will differ for each value of age.

The IRR Standard Error for the interactions are determined by first calculating the variance. The formula is:

$$V_{BxC} = V(\beta_1) + x^2V(\beta_3) + 2xCov(\beta_1,\beta_3)$$

In order to obtain the values necessary to calculate the variance, and therefore standard error, of the interaction, we need both the table of coefficients and the model variance-covariance matrix. I list both below:

PARAMETER ESTIMATES; COEFFICIENTS

docvis	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
outwork	.5195731	.014415	36.04	0.000	.4913202	.5478259
female	.3854858	.0110459	34.90	0.000	.3638362	.4071354
oxf	-.4545693	.0179558	-25.32	0.000	-.489762	-.4193767
married	-.1895326	.0367851	-5.15	0.000	-.2616301	-.1174351
age	.0180964	.0006517	27.77	0.000	.0168191	.0193736
mxa	.0030823	.0007826	3.94	0.000	.0015485	.0046161
_cons	.1022505	.0301124	3.40	0.001	.0432313	.1612697

VARIANCE-COVARIANCE MATRIX

```
. corr, _coef c             /// _coef c  = request variance-covariance matrix

             | docvis:
             |  outwork   female      oxf  married      age      mxa    _cons
-------------+------------------------------------------------------------------
docvis       |
     outwork |  .000208
      female |   .00005  .000122
         oxf | -.000203 -.000122  .000322
     married |  .000074 -6.3e-07 -.000094  .001353
         age | -1.3e-07 -2.0e-07 -8.3e-07  .000019  4.2e-07
         mxa | -1.4e-06  3.0e-07  1.6e-06 -.000028 -4.2e-07  6.1e-07
       _cons | -.000054 -.000052  .000104 -.000895 -.000018  .000018  .000907
```

$$V_{BxC} = V(\beta_1) + x^2V(\beta_3) + 2xCov(\beta_1,\beta_3)$$

$$V_{BxC} = .001353 + age^2*.00000061+ 2*age*-.000028$$

VARIANCE AND STANDARD ERROR OF INTERACTIONS

AGE = 30

```
. di .001353 + 30^2*.00000061+ 2*30*-.000028
.000222

. di sqrt(.000222)
.01489966               SE at age=30
```

AGE = 40

```
. di .001353 + 40^2*.00000061+ 2*40*-.000028
.000089

. di sqrt(.000089)
.00943398               SE at age=40
```

AGE = 50

```
. di .001353 + 50^2*.00000061+ 2*50*-.000028
.000078

. di sqrt(.000078)
.00883176               SE at age=50
```

AGE = 60

```
. di .001353 + 60^2*.00000061+ 2*60*-.000028
.000189

. di sqrt(.000189)
.01374773               SE at age=60
```

CONFIDENCE INTERVALS

Finally the confidence intervals must be calculated: The formula is

$$[\beta_1 + \beta_3 x] +/- z_{1-\alpha/2} * SE$$

I shall show the confidence intervals for the *mxa* interaction when age=30. Other confidence intervals can be calculated using the same logic. I use the 95% CI, or p=0.05. This statistic is calculated as 1.96.

LOW CI for AGE=30
. di (-.1895326 + .0030823*30) - 1.96*.01489966
-.12626693

. di exp(-.12626693)
.88137955 LOW CI for IRR of 0.9075

HIGH CI for AGE=30

. di (-.1895326 + .0030823*30) + 1.96*.01489966
-.06786027

. di exp(-.06786027)
.93439103 HIGH CI for IRR of 0.9075

We would report the interaction at age 30 as:
0.9075 (0.8814, 0.9344)

SIGNIFICANCE OF INTERACTION

The significance of the exponentiated interaction can be determined if the confidence interval includes 1.0. If it does, the predictor is not significant. Likewise, if the interaction is left un-exponentiated and the confidence interval includes 0, it is not significant. Since the confidence interval for the IRR of *mxa* at age 30 does not include 1.0, it is significant. at the p=0.05 level.

Another way to evaluate the significance of the interaction at a given value of the continuous predictor is to divide the coefficient by the SE. Doing so produces the Z or Wald statistic --- how you refer to it is based on the statistic used on it to calculate the p-value. For the Z-statistic it is assumed that it is distributed normally. However, as a 2-sided test you must multiply the resultant value by 2. If you use the Wald statistic, as in

SAS, then use the squared t-statistic to determine the probability. The probability of the interaction having a Z statistic as large as -6.51 for the *mxa* interaction with age=30 can be calculated as:

```
. di  (-.1895326 + .0030823*30)/ 0.01489966
-6.5144842

. di normprob(-6.5144842)*2
7.294e-11                          = 0.00000000007294
```

Or: 0.000

CATEGORICAL X CONTINUOUS INTERACTIONS

The same logic obtains for **Categorical X Continuous** interactions as it does for Binary X Continuous interactions, except that there are levels of binary variables each of which refers to a single reference level.

As discussed in the previous section, the formula used for a Binary X Continuous interaction is:

$$\text{IRR}_{\text{BxC}} = \exp[\beta_1 + \beta_3 x]$$

For the interaction effect where we do not parameterize the estimate as an incidence rate ratio, but as a simple coefficient, we drop the exponentiation:

$$\beta_{\text{BxC}} = \beta_1 + \beta_3 x$$

A binary variable is a categorical predictor with only two levels, 1 and 0. 0 is typically taken as the reference level. Therefore, for example, if we have an IRR of the number of fish caught based on the gender of the fish, *gender* is a binary predictor with, let's say, 1=male and 0=female. We compare the risk of male to female fish caught.

We generally refer to a variable as categorical when there are more than two levels of values, but at the same time we do not regard the variable as continuous. An example categorical variable is type of hospital admission: 1=emergency, 2=urgent, 3=elective.

When a categorical predictor is modeled, most statisticians factor or level the variable by creating as many binary or indicator variables as there are levels in the variable. Each new binary variable has the value of 1/0. 1 indicates that the subject or item has or is characterized by that level.

We factor the variable *type*, as defined above

```
. tab type, gen(type)

    TYPE OF |
  ADMISSION |      Freq.     Percent        Cum.
------------+-----------------------------------
      Emerg |      1,005       27.04       27.04
     Urgent |      1,307       35.16       62.20
   Elective |      1,405       37.80      100.00
------------+-----------------------------------
      Total |      3,717      100.00
```

New variables are created:
type1 = emergency 1=emergency; 0=not
type2 = urgent 1=urgent; 0=not
type3 = elective 1=elective; 0=not

If we wish to create an interaction of *type* and the continuous variable *age*, we must create the following two predictors.

. gen byte t2xage = type2*age
. gen byte t3xage = type3*age

Those observations in the model that do not have wither *t2xage*=1 or *t3xage*=1 are by default given the reference level of *t1xage*, which does not actually have to be created.

A model with length of stay, *los,* as the count response, and *age, type* and its interaction as the predictors appears as:

β_0 = intercept β_1 = age β_2 = type2
β_3 = type3 β_4 = t2xage β_5 = t3xage

I use the **poisson** command to model the data, but the **glm** command could have been used equally well.

POISSON MODEL WITH INTERACTION TERMS

```
. poisson los age type2 type3 t2xage t3xage, nolog

Poisson regression                                Number of obs   =       3717
                                                  LR chi2(5)      =    1193.02
                                                  Prob > chi2     =     0.0000
Log likelihood = -15148.244                       Pseudo R2       =     0.0379
------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0094958    .001649     5.76   0.000     .0062639    .0127278
       type2 |  -.2608531   .1801578    -1.45   0.148    -.6139559    .0922497
       type3 |  -.4714828   .1870624    -2.52   0.012    -.8381185   -.1048472
      t2xage |  -.0007527   .0024631    -0.31   0.760    -.0055803    .0040749
      t3xage |   .0005963   .0025697     0.23   0.817    -.0044403    .0056328
       _cons |   1.752462   .1210563    14.48   0.000     1.515196    1.989728
------------------------------------------------------------------------------
```

In terms of IRR, we have the following table, but without command and header statistics.

los	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.009541	.0016647	5.76	0.000	1.006284	1.012809
type2	.7703941	.1387925	-1.45	0.148	.5412057	1.096639
type3	.6240762	.1167412	-2.52	0.012	.4325236	.9004621
t2xage	.9992476	.0024612	-0.31	0.760	.9944353	1.004083
t3xage	1.000596	.0025712	0.23	0.817	.9955696	1.005649

INTERPRETATION

In a similar manner to Binary X Continuous interactions, the coefficient of each interaction is not the value used directly in interpretation. We calculate the interaction values for each non-reference level of the categorical predictor:

t2xage : $\beta_{\text{CatxCon2}} = \beta_2 + \beta_4 x$

$$\beta_{\text{CatxCon2}} = -0.2608531 - 0.0007527 * \text{age}$$

t3xage : $\beta_{\text{CatxCon3}} = \beta_3 + \beta_5 x$

$$\beta_{\text{CatxCon3}} = -0.4714828 + 0.0005963* \text{age}$$

In order to obtain the Incidence Rate Ratio, simply exponentiate the above formulae.

For *t2xage*, the IRR can be calculated for *age* 65 by:

AGE 65
. di exp(-0.2608531 - 0.0007527 * 65)
.73360936

IRR values for t2xage and t3xage can be shown as:

AGE	t2xage	t3xage
65	.734	.649
70	.731	.651
75	.728	.653
80	.725	.655
85	.723	.657
90	.720	.658

The IRR is *t2xage* at age 65 can be expressed as:

> Urgent admission patients at age 65 increase the probability or likelihood of the staying in the hospital 0.734 longer than elective patients. That is, age 65 elective admissions stay in the hospital some 27% longer than similar urgent admission patients.

STANDARD ERROR AND CONFIDENCE INTERVALS

The standard errors and confidence intervals of the above interactions are calculated exactly as they were for the Binary X Continuous interactions.

NOTE: The above methods are valid for use with both the Poisson and negative binomial models. With slight amendments, the methods also hold for logistic models. Care must be taken when both constructing and then interpreting interactions. Without a doubt they add complexity to a statistical model, but interactions often play an important role in model building.