APPLIED STATISTICS FOR SOCIAL SCIENTISTS (SOCIOLOGY)

WEEK 7: CATEGORICAL DEPENDENT VARIABLES 1

Topic: Contingency tables. Measures of association and independence. Tests for independence. Logistic regression models.

STATA commands and features: `tab hi2, logit`

Data set: BHPS 2001 (same as last week).

Readings: A. Agresti and B. Finlay (1997). [CHAPTER 15].

T.H. Wonnacott and R.J Wonnacott. 1990. [CHAPTER 17].

INTRODUCTION

Many if not most outcomes that we as sociologists are interested in are categorical rather than continuous. Even when we assume a continuous latent trait, we often (have to) use discrete measurements. Today we deal with the first class of outcomes, those that are categorical in nature and measurement.

1. CROSS-TABULATIONS AND TEST OF STATISTICAL INDEPENDENCE

In what follows, we examine how the proportions of smokers varies by social class and education. First we recode the variable *smoker* from 1=yes 2=no to a proper dummy. Then we ask for a crosstab.

```
recode smoker 1=1 2=0
tab msclass smoker, row
```

Social class - most recent job	smoker 0	yes	Total
professionals	242	33	275
	88.00	12.00	100.00
managers	1,449	316	1,765
	82.10	17.90	100.00
skilled workers	1,862	627	2,489
	74.81	25.19	100.00
semi/unskilled	833	401	1,234
	67.50	32.50	100.00
never had a job/armed	76	11	87
	87.36	12.64	100.00
Total	4,462	1,388	5,850
	76.27	23.73	100.00

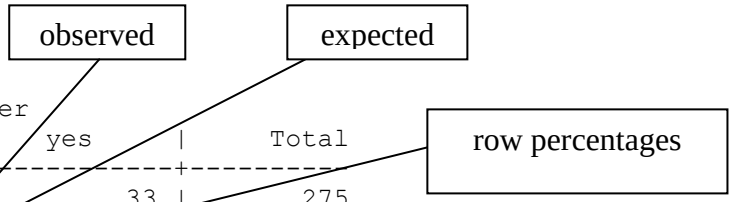
This table suggests that there is an association between smoking and social class. Is this association statistically significant?

We can run a test of the null hypothesis that the two variables are statistically independent. Suppose that there is no statistical association between the two variables, the rows of the table will all show the same proportions as the bottom row, known as the "marginal". The ***Chi-squared statistic*** which tests for independence is calculated as the difference between this hypothetical "independent" table, and the actual one.

To illustrate how it works, let's first ask *STATA* to print the expected frequencies and row frequencies in the cells of our table. We exclude msclass 5 ("never had a job/armed forces").

```
tab msclass smoker if msclass!=5, row exp
```

We get this table:



Social class - most recent job	smoker 0	yes	Total
professionals	242	33	275
	209.3	65.7	275.0
	88.00	12.00	100.00
managers	1,449	316	1,765
	1,343.3	421.7	1,765.0
	82.10	17.90	100.00
skilled workers	1,862	627	2,489
	1,894.3	594.7	2,489.0
	74.81	25.19	100.00
semi/unskilled	833	401	1,234
	939.2	294.8	1,234.0
	67.50	32.50	100.00
Total	4,386	1,377	5,763
	4,386.0	1,377.0	5,763.0
	76.11	23.89	100.00

How is the expected frequency calculated? If we assume there is no association between class and smoking, how many non-smokers would we expect in the first cell (class=1 professionals, smoker=0 no)? There are 275 respondents in this class. We would expect the percentage of non-smokers in this group of 275 to be equal to that in the marginal, in other words, equal to the percentage of non-smokers in the whole sample. Thus, we expect that 76.1% of the professionals are non-smokers: 0.761*275=209.3 (another way to write this is: row total x column total / overall total = 4386 * 275 / 5763 = 209.3). Likewise, among managers we would expect 0.761*2489=1894.3 non-smokers.

If we calculate the difference between the expected frequencies and the observed frequencies, we get an idea of how well the observed data fit a pattern of independence. We calculate the square of the deviation of the observed frequency from the expected frequency divided by the expected frequency (so, for the first cell above we get: $(\text{observed-expected})^2/\text{expected}=(242-209.3)^2/209.3=5.1$). The larger the

sum of these deviations, the less likely it is that there is no association between the two variables in the population from which we drew the sample.

The sum of deviations is called Pearson's χ^2 (chi-squared). In our example, adding all cells together gives us a χ^2 of 108.7. Can we reject the null-hypothesis that there is independence in the cross-table (no association)? To find out, we have to determine the probability of obtaining this value of χ^2 if the rows and columns were actually independent. We need the degrees of freedom to read this probability from a table with the χ^2 distribution. Luckily *STATA* does this automatically for us. The degrees of freedom in a cross-table are equal to: (number of rows - 1) * (number of columns - 1). In our case: (4-1)*(2-1)=3. The probability that we would observe Pearson's χ^2 statistic to be 108 or larger when we have 3 degrees of freedom is smaller than 0.001. We typically reject the null-hypothesis of independence when the *p*-value is less than 0.05 (or 5%).

In *STATA*, we get the χ^2 test like this:

```
tab msclass smoker if msclass!=5, nofreq chi
          Pearson chi2(3) = 108.7359   Pr = 0.000
```

We conclude that there is a statistically significant association between social class and smoking. The test doesn't indicate what this association looks like. We need to inspect the cross-table to get an idea of the nature of the association.

The χ^2 test does not tell us much about the "strength" of the relationship. With large samples we often find some departure from independence between any two categorical variables, even when the difference it quite small substantively. Below we will discuss more powerful and revealing forms of analysis.

EXERCISE 1

1. Following the example above, check whether there is an association between being employed (versus non-employed) and having a child in the household (you can use the information *nchild*).

2. Is the relationship between employment and children stronger or weaker for men than for women? Which statistic(s) could you use to illustrate the difference?

2. PROBABILITIES, ODDS, AND ODDS RATIOS

The odds of doing *X* rather than *Y* (or the odds of *X* over *Y*) is the probability of *X* divided by the probability of *Y*. For example, the odds of being a smoker:

$$\text{Odds of smoking} = \frac{\Pr(smoking)}{\Pr(notsmoking)}$$

Or, since the two probabilities sum to unity, we can write:

$$\text{Odds of smoking} = \frac{\Pr(smoking)}{1 - \Pr(smoking)}$$

To compare the odds for two different groups, A and B, we take the ratio of the odds. The value of the odds ratio (OR) depends on whether you have the odds for A divided by the odds for B or the other way round. Similarly, depending on whether you look at the odds of *X* over *Y* or the odds of *Y* over *X* you will get different answers. If you swap the order of A and B or the order of *X* and *Y* you will get the reciprocal of the previous odds ratio (*i.e.* 1/(OR)). This is why we say an odds ratio of 2 is equivalent in magnitude to an odds ratio of ½.

For example, the odds ratio of being a smoker for respondents in the semi/unskilled class compared to people in the professional class can be derived from the following table:

```
tab sclass smoker if (sclass==1 | sclass==4)
```

Social class - present job	smoker 0	yes	Total
professionals	191	27	218
semi/unskilled	382	201	583
Total	573	228	801

Odds of being a smoker in professional class = 27 / 191 = 0.141
Odds of being a smoker in semi/unskilled class = 201 / 382 = 0.526
OR of smoking in semi/unskilled vs prof class = 0.526/0.141=3.73

The odds ratio of 3.73 means that the odds of smoking for people in class 4 is almost 4 times higher than the odds for people in class 1. Or you can say that people in the professional class are 0.27 times as likely to be smokers as people in the semi/unskilled class (0.141/0.526 = 0.27 = 1/3.73).

SOME USEFUL PROPERTIES OF ODDS RATIO

1. Cross-product ratio
   In a 2x2 cross-table with cells A B
   C D
   The odds ratio (OR) = the ratio of the products of cell counts from diagonally opposite cells

$$OddsRatio = \frac{A \times D}{C \times B}$$

2. If the categories of one variable are switched the odds ratio in the new re-arranged table will equal 1/OR

3. When there is no association, OR = 1. Therefore, values of OR farther from 1 in a given direction represent stronger associations.

EXERCISE 2

Following Exercise 1 and the above example, what is the odds ratio of employment when we compare people under the age 60 with and without children in the household?
Is this association similar for men and women?

3. LOGISTIC REGRESSION: FROM LOG ODDS TO PROBABILITIES

We can model binary (and eventually other sorts of categorical) dependent variables using a technique known as logistic regression. There are various alternative ways of organising logistic regression results, some using odds ratios, but the approach we take here uses log odds (or "logits") instead. Here is what the log odds of smoking look like:

$$\log \text{odds of smoking} = \log \frac{\Pr(smoking)}{1 - \Pr(smoking)} \qquad \text{(Equation 1)}$$

Let's look at the log odds of smoking for people with primary or secondary education on the one hand and those with tertiary on the other hand. The crosstab looks like this:

```
gen primsec=hiqual>=3 if hiqual!=.
label var primsec "primary or secondary education"
tab primsec smoker, row chi
```

```
primary or |
 secondary |
 education |       smoker
1=yes 0=no |          0    yes      |     Total
-----------+----------------------+----------
         0 |      1,967        476 |     2,443
           |      80.52      19.48 |    100.00
-----------+----------------------+----------
         1 |      2,520        917 |     3,437
           |      73.32      26.68 |    100.00
-----------+----------------------+----------
     Total |      4,487      1,393 |     5,880
           |      76.31      23.69 |    100.00

          Pearson chi2(1) =  40.9031   Pr = 0.000
```

From this we can calculate the odds and odds ratio.

Now we run a logistic regression to obtain the log odds.

```
logit smoker primsec
```

We get this table:

```
Logistic regression                                     Number of obs   =       5880
                                                        LR chi2(1)      =      41.53
                                                        Prob > chi2     =     0.0000
Log likelihood = -3198.4518                             Pseudo R2       =     0.0065

------------------------------------------------------------------------------
      smoker |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     primsec |   .4079403   .0640044     6.37   0.000     .2824939    .5333866
       _cons |  -1.418847   .0510806   -27.78   0.000    -1.518963   -1.318731
------------------------------------------------------------------------------
```

The logistic regression model is based on *Maximum Likelihood Estimation*. It is a process of estimating the parameters (b_i) of a model so that the data are more likely to be observed than in the case of any other values of b_i (i.e. the log likelihood function is maximized). The *log likelihood* of the above logit model is a value that the computer tries to maximize by generating different sets of parameters for the model. It is therefore a statistic to show the *fit* of the model. The chi-squared test at the upper right corner tests the null hypothesis that the parameter(s) of the independent variable in the model equal zero.

The log odds of a person with primary or secondary education being a smoker is 0.408-1.419. Thus, the odds are: exp(0.408-1.419). The odds for a person with tertiary education is exp(-1.419). The ***odds ratio*** is the former divided by the latter. Or, such is the magic of logarithms, simply exp(0.408). To convert to an odds ratio, we simply exponentiate the relevant logit coefficient. *STATA* does this for us if we ask for logistic instead of logit output:

```
. logistic smoker primsec

Logistic regression                                     Number of obs   =       5880
                                                        LR chi2(1)      =      41.53
                                                        Prob > chi2     =     0.0000
Log likelihood = -3198.4518                             Pseudo R2       =     0.0065

------------------------------------------------------------------------------
      smoker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     primsec |   1.503717   .0962446     6.37   0.000     1.326434    1.704696
       _cons |   .2419929   .0123611   -27.78   0.000     .2189388    .2674746
------------------------------------------------------------------------------
```

Now, consider ***probabilities***. By rearranging the terms of Equation 1, we obtain:

$$\Pr(smoker) = \frac{1}{1+\exp(-\log\text{odds of smoker})} \qquad \textbf{(Equation 2)}$$

This logistic regression estimates the probability that people with the two different levels of educational qualification will be smokers or not. We can save the predicted probabilities by the following commands:

```
gen plogsmoker = _b[_cons] +_b[primsec]*primsec
gen psmoker= 1/(1+exp(-prlogsmoker))
```

Or, simply use the predict command:

```
predict psmoker
tabstat psmoker, by(primsec)
```

```
. tabstat psmoker, by(primsec)

Summary for variables: psmoker
     by categories of: primsec (primary or secondary
education 1=yes 0=no)

 primsec |      mean
---------+----------
       0 |  .1948424
       1 |  .2668024
---------+----------
   Total |  .2369048
--------------------
```

The variable *psmoker* takes one of just two values, depending on whether or not the respondent has a tertiary education. Note that the mean probabilities in this are the same as the percentages on page 5. The logit regression is to a certain extent a representation of the cross-tabulation, in a way that summarises the effect of the independent variable on the dependent one.

EXERCISE 3

Is there a (significant) association between age and voting *conservatives*? Demonstrate how the probability of voting *conservatives* varies with age by using **both** cross-tabulation **and** logistic regression. Construct 5-year age groups.

BONUS

Is the age pattern you found in Exercise 3 similar for home owners and people who do not own their homes? Compare the age pattern in the probabilities for these two groups.

4. CONCLUDING REMARKS

Odds and probabilities can be straightforwardly calculated from each other. They have their respective advantages. We usually start to think about probabilities, which provide rather straightforward and appropriate descriptions of changes and trends. The odds ratio has some nice properties too. For instance, you can multiply any row or column in a table by a non-zero positive number and the odds ratios will not change. In other words, odds ratios are not sensitive to the marginal distribution. Therefore, they allow us to consider historical change in degree of associations among such variables as educational qualifications, marital status, class or health while in effect controlling for the change in the absolute levels of any of these variables.