Universität Konstanz

Log-linear models for cross-tabulations using Stata

Maarten Buis

19th UK Stata Users Group meeting, 10 Sept. 2015

Outline

clarification of terminology

- Especially in economics the term log-linear models means
  - log transform the explained/dependent/left-hand-side/y-variable, and then
  - estimate a linear model using this transformed variable
- This is **not** how I will use that term
- Log-linear models is a set of models used to describe and test patterns in a cross-tabulation with 2 or more dimensions
- A useful analogy is that log-linear models are like ANOVA for categorical (ordinal) dependent variables.

What log-linear models are used for

- Log-linear models is a class of models that is used a lot in sociology
- A typical use would involve a table of the occupational class of the father against the occupational class of the son
- The two are related, but some cells need special attention
- For example, farmers mainly become farmers by inheriting a farm
- Log-linear models are used to quantify the association while still incorporating these special features.
- Such a flexible way of modeling cross tabulation is not only useful to sociologist, but a terminology has that proofed to be more of a hinderance.

An example: A 2 $\times$ 2 cross-tabulation

- The simplest cross-tabulation is a 2 by 2 table.
- Consider this German data from the ALLBUS (the German GSS) after reunification.

```
. tab east husb_career

           |  wife should support
 region of |   husband´s career
 residence |  disagree      agree |     Total
-----------+----------------------+----------
      west |     9,297      4,403 |    13,700 
      east |     5,639      1,770 |     7,409 
-----------+----------------------+----------
     Total |    14,936      6,173 |    21,109 
```

- This is easier to interpret with row percentages:

```
. tab east husb_career, row nofreq

           |  wife should support
 region of |   husband´s career
 residence |  disagree      agree |     Total
-----------+----------------------+----------
      west |     67.86      32.14 |    100.00 
      east |     76.11      23.89 |    100.00 
-----------+----------------------+----------
     Total |     70.76      29.24 |    100.00 
```

An example: Independence in a 2 × 2 cross-tabulation

- Remember the Pearson χ^2 test: $\chi^2 = \sum \frac{(O-E)^2}{E}$,
- where O are the observed counts and E are the expected counts if the variables are independent
- With independence we take the margins as given and distribute the observations over the cells such that there is no additional structure
- We know that $\frac{13{,}700}{21{,}109} \times 100\% = 64.90\%$ of the observations are from the west, and that overall $\frac{14{,}936}{21{,}109} \times 100\% = 70.76\%$ disagree
- So the expected count under independence for the West Germans who disagree is $0.6490 \times 0.7076 \times 21,109 = 9694$

```
. tab east husb_career, exp chi2 nokey

           |  wife should support
 region of |   husband´s career
 residence |  disagree      agree |     Total
-----------+----------------------+----------
      west |     9,297      4,403 |    13,700 
           |   9,693.6    4,006.4 |  13,700.0 
-----------+----------------------+----------
      east |     5,639      1,770 |     7,409 
           |   5,242.4    2,166.6 |   7,409.0 
-----------+----------------------+----------
     Total |    14,936      6,173 |    21,109 
           |  14,936.0    6,173.0 |  21,109.0 

          Pearson chi2(1) = 158.1252   Pr = 0.000
```

- We can reject the hypothesis that the two variables are independent

Independence and odds ratios

- Independence is one of the patterns in a cross-tabulation which can be tested with log-linear models.
- Such patterns are often framed as odds ratios
- An odds is the expected number of 'successes' per 'failure', and an odds ratio is a ratio of odds

		wives should support husband's career		
		disagree	agree	total
region of	west	9,694	4,006	13,700
residence	east	5,242	2,167	7,409
total		14,936	6,173	21,109

- So under independence the odds of agreeing for someone from the West is $\frac{4{,}006}{9{,}694} = .41$ or about two persons that agree for every five that disagree
- Under independence the odds of agreeing for someone from the East is $\frac{2{,}167}{5242} = .41$
- Independence means that the odds are the same, or their ratio is 1.

prepare the data

- The first step is to load the table as data in Stata
- If you start with individual level data, than `contract` is very useful.

```
. contract east husb_career, nomiss

. list

     +-----------------------------+
     | husb_c~r    east    _freq   |
     |-----------------------------|
  1. | disagree    west     9297   |
  2. |    agree    west     4403   |
  3. | disagree    east     5639   |
  4. |    agree    east     1770   |
     +-----------------------------+
```

estimate the independence model

- We can use `poisson` to estimate a model on these counts

```
. poisson _freq i.east i.husb_career, irr nolog

Poisson regression                              Number of obs     =          4
                                                LR chi2(2)        =    5653.89
                                                Prob > chi2       =     0.0000
Log likelihood = -101.13464                     Pseudo R2         =     0.9655

------------------------------------------------------------------------------
       _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        east |
       east  |   .5408029   .0077989   -42.63   0.000     .5257314    .5563065
             |
 husb_career |
      agree  |   .4132967   .0062536   -58.40   0.000     .4012199    .4257371
       _cons |   9693.647   93.26673   954.04   0.000     9512.561    9878.181
------------------------------------------------------------------------------

. est store indep
```

- The constant is the expected number of observations who are from the west and don't agree (both reference categories)
- The coefficient of `1.east` is the ratio by which this count increases/decreases when someone is from the east, i.e. it is the odds of coming from the east.
- The coefficient of `1.husb_career` is the odds of agreeing, which corresponds with the odds under independence we computed earlier.
- If we had included an interaction effect between `east` and `husb_career`, then that would represent the ratio of the odds of agreeing for West- and East-Germans, i.e. the odds ratio.
- By excluding that interaction we constrained the odds ratio to be 1

check if it is really an independence model

```
. predict mu
(option n assumed; predicted number of events)

. tabdisp east husb_career, cell(mu)

----------------------------------
             | wife should support
region of    |  husband´s career
residence    | disagree     agree
-------------+--------------------
        west | 9693.647  4006.353
        east | 5242.353  2166.647
----------------------------------

. tab east husb_career [fw=_freq], exp nofreq

             |  wife should support
  region of  |   husband´s career
  residence  |  disagree      agree |     Total
-------------+----------------------+----------
       west  |   9,693.6    4,006.4 |  13,700.0
       east  |   5,242.4    2,166.6 |   7,409.0
-------------+----------------------+----------
      Total  |  14,936.0    6,173.0 |  21,109.0
```

A likelihood ratio test for the independence model

- We can relax the independence assumption by adding an interaction effect between `east` and `husb_career`.

```
. poisson _freq i.east##i.husb_career, irr nolog

Poisson regression                              Number of obs     =          4
                                                LR chi2(3)        =    5815.16
                                                Prob > chi2       =     0.0000
Log likelihood = -20.497687                     Pseudo R2         =     0.9930

-----------------------------------------------------------------------------------
            _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
             east |
            east  |   .6065397   .0102377   -29.62   0.000     .5868024    .6269409
                  |
      husb_career |
           agree  |   .4735936    .008664   -40.85   0.000     .4569133    .4908829
                  |
 east#husb_career |
      east#agree  |   .6627738   .0217506   -12.53   0.000     .6214855     .706805
                  |
            _cons |       9297   96.42095   881.04   0.000     9109.926    9487.915
-----------------------------------------------------------------------------------

. est store sat

. lrtest indep sat

Likelihood-ratio test                                 LR chi2(1)  =    161.27
(Assumption: indep nested in sat)                     Prob > chi2 =    0.0000
```

- This interaction effect is the odds ratio.
- The odds of agreeing in the East is .66 times the odds of agreeing in the West.
- The odds of agreeing in the East is (.66-1)*100%= -34% less than the odds of agreeing in the West.
- Not surprisingly this difference is statistically significant.

Log-linear models for a $2 \times 2 \times 2$ table

- This difference could be the result of the fact that the female labor force participation in the former GDR (East-Germany) was a lot higher than the FRG (West-Germany).
- Alternatively, the GDR was very effective at suppressing religion, and religious people were more likely to agree

```
. tab east relig, row nofreq

 region of | religious affiliation
 residence | no affili  an affili |     Total
-----------+----------------------+----------
      west |     12.53      87.47 |    100.00 
      east |     68.23      31.77 |    100.00 
-----------+----------------------+----------
     Total |     26.09      73.91 |    100.00 

. tab relig husb_career, row nofreq

               | wife should support
     religious |  husband´s career
   affiliation |  disagree      agree |     Total
---------------+----------------------+----------
no affiliation |     79.77      20.23 |    100.00 
an affiliation |     66.20      33.80 |    100.00 
---------------+----------------------+----------
         Total |     70.76      29.24 |    100.00 
```

- If the latter mechanism is the only reason, then the independence model should fit within the religious and non-religious sub-tables

prepare the data

```
. contract husb_career east relig, nomiss
. tabdisp east husb_career relig, cell(_freq)
```

region of residence	religious affiliation and wife should support husband´s career — no affiliation — disagree	agree	— an affiliation — disagree	agree
west	1572	386	7690	3998
east	4073	1046	1551	720

estimate the conditional independence model

```
. poisson _freq i.husb_career##i.relig i.east##i.relig, irr nolog

Poisson regression                              Number of obs     =          8
                                                LR chi2(5)        =   14883.91
                                                Prob > chi2       =     0.0000
Log likelihood = -40.136821                     Pseudo R2         =     0.9946

---------------------------------------------------------------------------------
                  _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------+--------------------------------------------------------
            husb_career |
                 agree  |   .2536758   .0075059   -46.36   0.000     .2393831     .268822
                        |
                  relig |
         an affiliation |   4.954244   .1281133    61.88   0.000     4.709404    5.211814
                        |
      husb_career#relig |
   agree#an affiliation |   2.012611   .0695921    20.23   0.000     1.880733    2.153737
                        |
                   east |
                  east  |   2.614402   .0694701    36.17   0.000     2.481728    2.754169
                        |
             east#relig |
    east#an affiliation |   .0743198   .0026086   -74.06   0.000      .069379    .0796125
                        |
                  _cons |   1561.807   36.51325   314.54   0.000     1491.857    1635.037
---------------------------------------------------------------------------------

. est store cindep
```

Does this model fit?

```
. predict mu
(option n assumed; predicted number of events)

. tabdisp east husb_career relig, cell(_freq mu) format(%9.0f)
```

region of residence	religious affiliation and wife should support husband´s career: no affiliation, disagree	no affiliation, agree	an affiliation, disagree	an affiliation, agree
west	1572	386	7690	3998
	1562	396	7738	3950
east	4073	1046	1551	720
	4083	1036	1503	768

Does this model fit? (2)

- A common way of summarizing the fit is the index of dissimilarity, the proportion of observations that need to be ‘shifted’ in order to fully fit the data

```
. sum _freq , meanonly
. local n = r(sum)
. gen d = abs(_freq/`n´-mu/`n´)
. sum d, meanonly
. di "index of dissimilarity = " r(sum)/2
index of dissimilarity = .00549226
```

- Alternatively, one can compare the model with the fully saturated model (the model with the best possible fit) using
    - a likelihood ratio test
    - BIC (negative values show support for the constrained model, positive values for the saturated model)

```
. qui poisson _freq i.husb_career##i.east##i.relig
. estimates store sat
. lrtest cindep sat
Likelihood-ratio test                                 LR chi2(2)  =      5.82
(Assumption: cindep nested in sat)                    Prob > chi2 =    0.0544
. di "BIC = " r(chi2) - r(df)*ln(`n´)
BIC = -14.086432
```

Compare with a model with an effect of `east`

```
. poisson _freq i.husb_career##i.east i.husb_career##i.relig i.east##i.relig, irr nolog
Poisson regression                              Number of obs     =          8
                                                LR chi2(6)        =   14886.05
                                                Prob > chi2       =     0.0000
Log likelihood = -39.067483                     Pseudo R2         =     0.9948

------------------------------------------------------------------------------------
                 _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------
           husb_career |
                 agree |    .264348   .0107801   -32.63   0.000     .2440418    .2863439
                       |
                  east |
                  east |   2.645171   .0734729    35.02   0.000     2.505016    2.793167
                       |
      husb_career#east |
            agree#east |   .9443658    .036985    -1.46   0.144     .8745888     1.01971
                       |
                 relig |
        an affiliation |   4.980792   .1304105    61.32   0.000      4.73164    5.243064
                       |
     husb_career#relig |
  agree#an affiliation |   1.949286   .0796728    16.33   0.000     1.799221    2.111867
                       |
            east#relig |
   east#an affiliation |   .0748718   .0026527   -73.16   0.000      .069849    .0802558
                       |
                 _cons |   1548.624   37.40564   304.09   0.000     1477.019    1623.701
------------------------------------------------------------------------------------

. est store east

. lrtest cindep east

Likelihood-ratio test                                 LR chi2(1)  =      2.14
(Assumption: cindep nested in east)                   Prob > chi2 =    0.1436
```

log-linear models and logit models

- We could also estimate this model with `logit`

```
. poisson _freq i.husb_career##relig i.east##relig, irr nolog
Poisson regression                              Number of obs     =          8
                                                LR chi2(5)        =   14883.91
                                                Prob > chi2       =     0.0000
Log likelihood = -40.136821                     Pseudo R2         =     0.9946

----------------------------------------------------------------------------------------
                 _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+----------------------------------------------------------------
           husb_career |
                 agree |   .2536758   .0075059   -46.36   0.000     .2393831     .268822
                       |
                 relig |
        an affiliation |   4.954244   .1281133    61.88   0.000     4.709404    5.211814
                       |
     husb_career#relig |
  agree#an affiliation |   2.012611   .0695921    20.23   0.000     1.880733    2.153737
                       |
                  east |
                  east |   2.614402   .0694701    36.17   0.000     2.481728    2.754169
                       |
            east#relig |
   east#an affiliation |   .0743198   .0026086   -74.06   0.000      .069379    .0796125
                       |
                 _cons |   1561.807   36.51325   314.54   0.000     1491.857    1635.037
----------------------------------------------------------------------------------------

. logit husb_career i.relig [fw=_freq], or nolog
Logistic regression                             Number of obs     =     21,036
                                                LR chi2(1)        =     434.48
                                                Prob > chi2       =     0.0000
Log likelihood = -12493.716                     Pseudo R2         =     0.0171

--------------------------------------------------------------------------------
   husb_career | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
         relig |
an affiliation |   2.012611   .0695921    20.23   0.000     1.880732    2.153737
         _cons |   .2536759   .0075059   -46.36   0.000     .2393831     .268822
--------------------------------------------------------------------------------
```

Notation for models

- It is customary to refer to the models using a short hand like [RW][ER]
- The letters are abbreviations for variables

  E `east`

  W `husb_career`

  R `relig`

- letters grouped together are variables grouped together in Stata's factor variable notation with the #

notation	factor variable notation
[W][E][R]	`i.husb_career i.east i.relig`
[RW][ER]	`i.relig##i.husb_career i.east##i.relig`
[EW][WR][ER]	`i.east##i.husb_career i.husb_career##i.relig i.east##i.relig`
[WER]	`i.husb_career##i.east##i.relig`

An example: homogamy

- We can look at the education of both partners, again using the German ALLBUS data

```
. tab meduc feduc, row nokey
```

male education	female education low	lower voc	medium vo	higher vo	universit	Total
low	2,068 61.20	703 20.80	426 12.61	122 3.61	60 1.78	3,379 100.00
lower voc.	4,555 30.52	7,200 48.25	2,523 16.91	416 2.79	229 1.53	14,923 100.00
medium voc.	1,032 11.38	1,792 19.76	4,845 53.42	856 9.44	544 6.00	9,069 100.00
higher voc.	334 9.45	472 13.36	1,157 32.74	1,100 31.13	471 13.33	3,534 100.00
university	389 6.15	740 11.69	1,783 28.17	999 15.78	2,418 38.21	6,329 100.00
Total	8,378 22.50	10,907 29.29	10,734 28.83	3,493 9.38	3,722 10.00	37,234 100.00

Compare the independent and saturated models

```
. contract meduc feduc, nomiss
. qui poisson _freq i.meduc##i.feduc, irr
. est store full
.
. qui poisson _freq i.meduc i.feduc, irr
. est store indep
. llingov , sat(full)
```

	LL	df	p	BIC	D
r1	17484.97	16	0	17316.57	.289634

What is `llingov`?

```
program define llingov, rclass
    syntax, sat(name)
    if "`e(cmd)'" != "poisson" {
        di as error "llingov only works after poisson"
        exit 198
    }
    // index of dissimilarity
    local y "`e(depvar)'"
    tempvar diff
    tempname res
    qui predict double `diff' if e(sample), n
    qui replace `diff' = abs(`y' - `diff')
    sum `y' if e(sample), meanonly
    local n = r(sum)
    sum `diff' if e(sample), meanonly
    local d = r(sum)/(2*`n')

    // likelihood ratio and BIC
    qui lrtest . `sat'
    local p = r(p)
    local df = r(df)
    local ll = r(chi2)
    local bic = r(chi2) - r(df)*ln(`n')
    matrix `res' = `ll', `df', `p', `bic', `d'
    matrix colname `res' = "LL" "df" "p" "BIC" "D"
    matlist `res'
    return matrix res `res'
end
```

Quasi-independence model

- Lets start with taking care of the diagonals
- We assume there are two groups:
  - there is a group that insist on someone with the same education
  - there is another group that randomly falls in love

```
. gen diag = (meduc==feduc)*meduc

. tabdisp meduc feduc, cell(diag)
```

male education	female education low	lower voc.	medium voc.	higher voc.	university
low	1	0	0	0	0
lower voc.	0	2	0	0	0
medium voc.	0	0	3	0	0
higher voc.	0	0	0	4	0
university	0	0	0	0	5

```
. label value diag ed
```

fit the quasi-independence model

```
. poisson _freq i.meduc i.feduc i.diag, irr nolog

Poisson regression                              Number of obs     =         25
                                                LR chi2(13)       =   31953.20
                                                Prob > chi2       =     0.0000
Log likelihood = -2548.7891                     Pseudo R2         =     0.8624

------------------------------------------------------------------------------
       _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       meduc |
 lower voc.  |   5.730713   .1767324    56.61   0.000     5.394584    6.087785
medium voc.  |   3.406199   .1108968    37.64   0.000     3.195634    3.630637
higher voc.  |   1.516249   .0525283    12.01   0.000     1.416713    1.622778
 university  |   2.320418   .0748851    26.08   0.000     2.178191    2.471931
             |
       feduc |
 lower voc.  |   .9246445   .0203206    -3.56   0.000     .8856625    .9653422
medium voc.  |   1.145789   .0216409     7.21   0.000     1.104149    1.188999
higher voc.  |   .3949555   .0095705   -38.34   0.000     .3766361    .4141659
 university  |   .2300723   .0070539   -47.93   0.000     .2166542    .2443214
             |
        diag |
         low |   4.251879   .1642796    37.46   0.000     3.941787    4.586367
 lower voc.  |   2.793698   .0721779    39.76   0.000     2.655754    2.938807
medium voc.  |   2.552405   .0689381    34.69   0.000     2.420803    2.691161
higher voc.  |    3.77663   .1606407    31.24   0.000     3.474547    4.104977
 university  |   9.312283   .3617783    57.44   0.000     8.629534    10.04905
             |
       _cons |   486.3731   15.45148   194.75   0.000     457.0124    517.6202
------------------------------------------------------------------------------

. llingov, sat(full)

             |        LL          df           p         BIC           D
-------------+----------------------------------------------------------
          r1 |  4882.975          11           0    4767.201    .1155445
```

Interpret the coefficients

```
. predict mu, n

. tabdisp meduc feduc, c(mu)
```

male education	female education low	lower voc.	medium voc.	higher voc.	university
low	2068	449.7223	557.281	192.0957	111.901
lower voc.	2787.265	7200	3193.617	1100.846	641.2723
medium voc.	1656.683	1531.843	4845	654.3163	381.157
higher voc.	737.4627	681.8909	844.9767	1100	169.6697
university	1128.589	1043.544	1293.125	445.7424	2418

```
. di exp(_b[_cons]) * exp(_b[1.diag])
2068

. di exp(_b[_cons]) * exp(_b[2.meduc]) * exp(_b[2.feduc]) * exp(_b[2.diag])
7200

. di ( 681.8909 /  737.4627 ) / ( 1043.544 /  1128.589 )
.99999973
```

Adding a diagonal

- The fit was not very good, so lets assume there is a third group: those that move one step up or down

```
. gen move_sym = abs(feduc-meduc) == 1

. tabdisp meduc feduc, cell(move_sym)
```

male education	female education low	lower voc.	medium voc.	higher voc.	university
low	0	1	0	0	0
lower voc.	1	0	1	0	0
medium voc.	0	1	0	1	0
higher voc.	0	0	1	0	1
university	0	0	0	1	0

Fit the model

```
. poisson _freq i.meduc i.feduc i.diag i.move_sym, irr nolog

Poisson regression                              Number of obs     =         25
                                                LR chi2(14)       =   34910.26
                                                Prob > chi2       =     0.0000
Log likelihood = -1070.2624                     Pseudo R2         =     0.9422

------------------------------------------------------------------------------
       _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       meduc |
 lower voc.  |   3.338166   .1149113    35.02   0.000     3.120374     3.57116
medium voc.  |   2.707588   .0873711    30.87   0.000     2.541647    2.884363
higher voc.  |   1.230837   .0446596     5.72   0.000     1.146345    1.321555
 university  |   2.929866   .0981173    32.10   0.000     2.743735    3.128624
             |
       feduc |
 lower voc.  |   .6559253   .0166208   -16.64   0.000     .6241449    .6893239
medium voc.  |   .9065449    .017194    -5.17   0.000     .8734639    .9408787
higher voc.  |   .3061911   .0083201   -43.56   0.000     .2903107    .3229403
 university  |   .3186353    .010115   -36.03   0.000     .2994145      .33909
             |
        diag |
        low  |   5.285556   .2107696    41.75   0.000     4.888186    5.715229
 lower voc.  |    8.40447   .3063953    58.39   0.000     7.824899    9.026968
medium voc.  |   5.045011   .1533115    53.26   0.000     4.753299    5.354625
higher voc.  |   7.460018   .3595832    41.69   0.000     6.787514    8.199152
 university  |    6.61995   .2616735    47.82   0.000     6.126443    7.153211
             |
  1.move_sym |   2.773769   .0548879    51.56   0.000      2.66825    2.883461
       _cons |   391.2549    13.0152   179.45   0.000     366.5594    417.6142
------------------------------------------------------------------------------

. llingov, sat(full)
             |        LL          df           p         BIC           D
-------------+------------------------------------------------------------
          r1 |  1925.922          10           0    1820.672    .0590599
```

interpret the coefficients

```
. predict mu, n

. tabdisp meduc feduc, c(mu)
```

male education	female education low	lower voc.	medium voc.	higher voc.	university
low	2068	711.8434	354.6902	119.7988	124.6676
lower voc.	3622.747	7200	3284.183	399.9083	416.1613
medium voc.	1059.357	1927.379	4845	899.7156	337.5486
higher voc.	481.5709	315.8745	1210.932	1100	425.6224
university	1146.325	751.9033	1039.195	973.5774	2418

```
. di exp(_b[_cons]) * exp(_b[1.diag])
2068

. di exp(_b[_cons]) * exp(_b[2.meduc]) * exp(_b[1.move_sym])
3622.7474

. di exp(_b[_cons]) * exp(_b[3.meduc])
1059.3571

. di ( 315.8745 / 481.5709 ) / ( 751.9033 / 1146.325 )
1.0000002
```

Adding asymmetry

- descriptively we found that men were more likely to marry ‘down’ than ‘up’
- lets incorporate that in our previous model

```
. gen move_asym = (meduc-feduc==1) + 2*(meduc-feduc==-1)

. tabdisp meduc feduc, cell(move_asym)
```

male education	female education low	lower voc.	medium voc.	higher voc.	university
low	0	2	0	0	0
lower voc.	1	0	2	0	0
medium voc.	0	1	0	2	0
higher voc.	0	0	1	0	2
university	0	0	0	1	0

```
. label define m 1 "down" 2 "up"

. label value move_asym m
```

Fit the model

```
. poisson _freq i.meduc i.feduc i.diag i.move_asym, irr nolog

Poisson regression                              Number of obs     =         25
                                                LR chi2(15)       =   35202.87
                                                Prob > chi2       =     0.0000
Log likelihood = -923.95537                     Pseudo R2         =     0.9501

------------------------------------------------------------------------------
       _freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       meduc |
  lower voc. |   3.176795   .1089642    33.70   0.000     2.970249    3.397703
 medium voc. |   2.495323   .0816923    27.93   0.000     2.340238    2.660686
 higher voc. |   1.041134   .0389619     1.08   0.281     .9675032    1.120368
  university |   2.524221    .085721    27.27   0.000      2.36168    2.697948
             |
       feduc |
  lower voc. |   .6979436   .0180204   -13.93   0.000     .6635031    .7341717
 medium voc. |   1.104936   .0247197     4.46   0.000     1.057534    1.154464
 higher voc. |   .3500088   .0098091   -37.46   0.000     .3313019    .3697721
  university |   .3687206   .0121422   -30.30   0.000     .3456741    .3933037
             |
        diag |
         low |   5.169272   .2041837    41.59   0.000     4.784179    5.585364
  lower voc. |   8.117119   .2962936    57.37   0.000     7.556681    8.719122
 medium voc. |   4.392467   .1386041    46.90   0.000     4.129038    4.672703
 higher voc. |   7.545465   .3634368    41.96   0.000     6.865732    8.292495
  university |   6.493972   .2571544    47.25   0.000     6.009022    7.018061
             |
   move_asym |
        down |   3.057201   .0624071    54.74   0.000       2.9373    3.181997
          up |   2.082365   .0544486    28.05   0.000     1.978336    2.191864
             |
       _cons |   400.0563    13.1268   182.60   0.000     375.1381    426.6297
------------------------------------------------------------------------------

. llingov, sat(full)

             |        LL         df          p        BIC          D
-------------+------------------------------------------------------
          r1 |  1633.308          9          0   1538.583   .0581226
```

Interpret the coefficients

```
. predict mu, n

. tabdisp meduc feduc, c(mu)
```

male education	female education: low	lower voc.	medium voc.	higher voc.	university
low	2068	581.431	442.0367	140.0232	147.509
lower voc.	3885.387	7200	2924.181	444.8251	468.6058
medium voc.	998.2699	2130.063	4845	727.585	368.0827
higher voc.	416.5121	290.702	1406.983	1100	319.8025
university	1009.83	704.8046	1115.798	1080.567	2418

```
. di exp(_b[_cons]) * exp(_b[1.diag])
2068

. di exp(_b[_cons]) * exp(_b[2.meduc]) * exp(_b[1.move_asym])
3885.3876

. di exp(_b[_cons]) * exp(_b[2.feduc]) * exp(_b[2.move_asym])
581.431
```

Unidiff models

- This table involves respondents that were born between 1900 and 1993, we may want to adjust for that
- We could do that as before
- Alternatively, we could model the table for the oldest cohort and say that the next cohort is the same except that all the parameters are x percent larger or smaller
- So the pattern remains the same, but the strength of the association increases or decreases by x percent.
- You need a user written package to estimate that: `unidiff` by Maurizio Pisati

Estimation of a unidiff model

```
. unidiff _freq, row(meduc) col(feduc) layer(coh) ///
>         effect(mult) pattern(fi) lambda(rawlog)
  (output omitted )
Table structure
-------------------------------------------------------------------------------
            Name      Label                                  N. of categories
-------------------------------------------------------------------------------
Row         meduc     male education                                  5
Column      feduc     female education                                5
Layer       coh                                                       4
-------------------------------------------------------------------------------

Model specification
-------------------------------------------------------------------------------
Layer effect:            multiplicative
R-C association pattern: full interaction
Additional variables:    none
-------------------------------------------------------------------------------

Goodness-of-fit statistics
-------------------------------------------------------------------------------
Model                 N    df      X2     p        G2     p     rG2     BIC    DI
-------------------------------------------------------------------------------
Cond. indep.     37165   64 17778.3  0.00  15352.8  0.00    0.0  14679.3  26.1
Null effect      37165   48   254.6  0.00    247.7  0.00   98.4   -257.4   2.6
Multipl. effect  37165   45   239.5  0.00    237.1  0.00   98.5   -236.4   2.5
-------------------------------------------------------------------------------
```

Interpretation of a unidiff model

Phi parameters (layer scores)

coh	Raw	Scaled 1	Scaled 2
1900	2.7623	1.0000	0.5223
1925	2.6491	0.9590	0.5009
1950	2.7238	0.9861	0.5150
1975	2.4296	0.8796	0.4594

Psi parameters (R-C association scores)

male education	female education low	lower	medium	higher	univer
low	0.00	0.00	0.00	0.00	0.00
lower voc.	0.00	0.57	0.43	0.30	0.29
medium voc.	0.00	0.59	1.13	0.98	1.05
higher voc.	0.00	0.53	1.04	1.51	1.44
university	0.00	0.65	1.25	1.58	2.13

Interpretation of a unidiff model (2)

Total interaction effects (raw) - Additive form

coh and male education	female education low	lower	medium	higher	univer
1900					
low	0.00	0.00	0.00	0.00	0.00
lower voc.	0.00	1.59	1.19	0.82	0.81
medium voc.	0.00	1.64	3.11	2.70	2.90
higher voc.	0.00	1.46	2.87	4.16	3.97
university	0.00	1.80	3.45	4.36	5.87
1925					
low	0.00	0.00	0.00	0.00	0.00
lower voc.	0.00	1.52	1.14	0.78	0.78
medium voc.	0.00	1.58	2.99	2.59	2.78
higher voc.	0.00	1.40	2.75	3.99	3.81
university	0.00	1.73	3.31	4.18	5.63
1950					
low	0.00	0.00	0.00	0.00	0.00
lower voc.	0.00	1.56	1.17	0.81	0.80
medium voc.	0.00	1.62	3.07	2.66	2.86
higher voc.	0.00	1.44	2.83	4.10	3.91
university	0.00	1.78	3.40	4.30	5.79
1975					
low	0.00	0.00	0.00	0.00	0.00
lower voc.	0.00	1.40	1.05	0.72	0.71
medium voc.	0.00	1.44	2.74	2.38	2.55
higher voc.	0.00	1.29	2.52	3.66	3.49
university	0.00	1.58	3.03	3.83	5.17

```
. di 2.4296*.65
1.57924

. di 1.58/1.80
.87777778
```

Summary

- Log-linear models describe and test patterns in cross-tabulations
- The simplest pattern is independence, the counts in cells are only determined by the margins
- Many of these models can be estimated using `poisson`
- With higher dimensional tables we can look if independence holds within sub-tables
- A more complex model is quasi-independence. There are two groups: one stays on the diagonal and one follows a independence pattern
- We can complicate the model even more, for example by adding additional diagonals, but there are many more ways of describing such tables.
- We can compare tables by saying that the basic structure is the same, but all the effects are x% larger are smaller than the reference table.
- What I did not discuss are log-linear models for ordinal variables, common models for such tables are stereotyped ordered regression and the RCII (Row Column II) model.