I'm not robot
reCAPTCHA
Continue

Log linear model stata

Contingency tables are enough to analyze associations in a two-way table. Three-way tables and higher have more associations that may be explored requiring flexible methods to generate the expected frequency for different hypotheses. The loglinear regression model is one approach that can be used. Loglinear regression models have a common form We will use the ipf (iterated proportional fitting) command written by Adrian Mander to estimate the model. Use findit ipf in Stata 7 to find commands. Iterated proportional fittings are a fast and efficient method for estimating the expected frequency for various models. The downside to using the ipf command is that it does not produce a coefficient for each individual term in the model. This results in Possibility and Pearson chi-squares along with the expected frequency for each cell. We will illustrate the loglinear regression model for contingency tables using acm data sets from Agresti. The ACM file contains information about the use of alochol, cigarettes and cannabis. Let's start with a two-variable example looking at the relationship between smoking and marijuana use. We'll start with a traditional contingency table analysis of the independence between cigarettes and marijuana. Example Two Variables using clearly contains data from obs: 8 vars: 4 28 Nov 2001 size 14:28: 72 (100.0% of free memory) ---------- displays variable value variable format variable label format ---------- byte label %8.0g yn alcohol use c byte %8.0g yn cigarette use m byte %8.0g yn marijuana using freq int %8.0g ---------- list c m freq 1. no yes 2 2. Well yes yes 3 3. Well yes no 43 4. Yes no yes 44 5. no. 279 6. Yes no 456 7. Yes, there is no 538 8. yes yes yes 911 tabulate c m [fw=freq], exp all +----------+ | Key | |----------| | frequency | | expected frequency | +----------+ cigarettes | marijuana use | no yes | Total ----------+----------+---------- no | 735 46 | 781 | 451.6 329.4 | 781.0 ----------+----------+---------- yes | 581 914 | 1,495 | 864.4 630.6 | 1,495.0 ----------+----------+---------- Total | 1,316 960 | 2,276 | 1,316.0 960.0 | 2276.0 Pears chi2(1) = 642.0347 Pr = 0.000 chi2(1) probability ratio = 751.8083 Pr = 0.000 Cramer V = 0.5311 gamma = 0.9235 ASE = 0.012 Tau-b Kendall = 0.5311 ASE = 0.014 ipf [fw=freq], fit(c+m) exp Expansion of various marginal models marginal model 1 varlist: c marginal model 2 varlist : m unique varlist c m m N.B. structural/sampling zeroes may lead to an incorrect df Residual degrees of freedom = 1 Goodness of Fit Tests ---------- df = 1 Likelihood Ratio Statistic G^2 = 751.8083 p-value = 0.000 Pearson Statistic X^2 = 642.0347 p-value = 0.000 c m Efreq Ofreq prob 1 1 630.57996 914 .27705622 1 2 864.42004 581 .37979791 2 1 329.42004 46 .1447364 2 2 451.57996 735 .19840947 ipf [fw=freq], fit(c*m) exp Expansion of the various marginal models ---------- marginal model 1 varlist : c m unique varlist c m ---------- N.B. structural/sampling zeroes may lead to an incorrect df Residual degrees of freedom = 0 Goodness of Fit Tests ---------- df = 0 Likelihood Ratio Statistic G^2 = 0.0000 p-value = . Pearson Statistics X^2 = 0.0000 p-value = . c m Efreq Ofreq prob 1 1 914 914 .40158172 1 2 581 581 .25527241 2 1 46 46 .0202109 2 2 735 735 .32293497 Last model, the model, using c*m, reproduces the perfectly observed frequency because it uses information regarding the distribution of cigarettes and marijuana joints and marginal. This is equivalent to testing c+m+c*m and is called a saturated model because it is completely deterministic. The loglinear model for independence looks like this: The loglinear model for saturated cases looks like this: logμij = τττ + ττic + τjm + τijcm When you get to three variables and a higher model the question that can be asked is more interesting than just two variables. If we add alcohol to our model then testing a+c+m is equivalent to running a 3 way contingency table test for independence. It is known as the model of gotong royong independence. It's not really very interesting and thankfully not all that common either. The loglinear model for shared independence would look like this: logμijk = ττ + ττia + τjc + κkm The short notation for this model is (a,c,m). Consider the following model: logμijk = ττ + ττia + τjc + τkm + τikam It is a model of mutual independence that affirms two-way independence between cigarettes and four combinations of alcohol and marijuana content. Consider the following models of conditional independence: logμijk = ττ + ττia + ττjc + τκkm + τijcm + τττikam This model contributes to the relationship between cigarettes and marijuana control for alochol and for the relationship between alcohol and marijuana control for cigarettes. It is a model of conditional independence that determines the conditional independence between alcohol and cigarettes while controlling marijuana. It can be marked (cm, am). Models that contain all two-way interactions are known as homogeneous association models. logμijk = τττ + τia + ττjc + κkm + ττijac + τjkcm In this model all three variable pairs are conditionally dependent on the third variable and (am, cm, ac). The saturated model, (acm), with all possible interactions, looks like this: logμijk = τττ + τττia + τjc + τkm + τijac + ττjkcm + τijkacm Three Variable Example Table c m [fw=freq], by (a) ---------- alcohol | usage and | cigarette marijuana | use use | no ya ---------- +---------- not | no | 279 2 yes | 43 3 ----------+---------- yes | no | 456 44 yes | 538 911 ---------- ipf [fw=freq], fit(a+c+m) /* shared independence model */ Expansion of various marginal models ---------- marginal model 1 varlist : varlist marginal model 2 : c marginal model 3 varlist : m varlist unique c m ---------- N.B. structural zero/sampling can cause incorrect DF Residual freedom level = 4 Goodness Fit Tests ---------- df = Statisti Probability Ratio 4 G^2 = 1286.0199 p-value = 0.000 Pearson Statistic X^2 = 1411.3860 p-value = 0.000 ipf [fw=freq], fit(a*c+m) Expansion of various marginal models ---------- marginal model 1 varlist: c marginal model 2 varlist : m varlist unique c m ---------- N.B. structural / sampling zero can cause incorrect residual DF levels freedom = 3 Goodness Fit Test ---------- df = 3 Possible Statistical Ratio G^2 = 843.8267 p-value = 0.000 Pearson Statistics X^2 = 704.9071 p-value = 0.000 ipf [fw=freq] , fit(a*m+c) Expansion of various marginal models ---------- marginal model 1 varlist: varlist m marginal model 2: c unique varlist m c ---------- N.B. structural zero/sampling can cause incorrect DF Residual freedom level = 3 Goodness Fit Tests ---------- df = Statistical Ratio 3 Probability G^2 = 939.5626 p-value = 0.000 Pearson Statistic X^ = 824.1630 p-value = 0.000 ipf [fw=freq], fit(c*m+a) Expansion of various marginal models ---------- marginal model 1 varlist: c m marginal model 2 varlist : unique varlist c m a ---------- N.B. structural / sampling zero can cause incorrect residual DF levels of freedom 3 Good Fit Tests ---------- df = 3 Possible Statistical Ratios G^2 = 534.2117 p-value = 0.000 Pearson Statistics X^2 = 505.5977 p-value = 0.000 ipf [fw=freq] , fit(c*m+a*m) Expansion of various marginal models ---------- model 1 varlist : c m marginal model 2 varlist : a m unique varlist c m a ---------- N.B. structural / sampling zero can cause wrong df Residual degree of freedom = 2 Goodness Fit ---------- df = 2 Possible Statistical Ratio G^2 = 187.7543 p-value = 0.000 Pearson Statistics X^2 = 177.6149 p-value = 0.000 ipf [fw=freq], fit(c*a+m*a) Expansion of various marginal ---------- models of model 1 varlist: c marginal model 2 varlist : m varlist c a m ---------- N.B. structural zero/sampling can cause incorrect DF Residual freedom level = 2 Goodness Fit Tests ---------- df = 2 Possible Statistical Ratio G^2 = 497.3693 p-value = 0.0 Pearson Statistic X^2 = 443.7611 p-value = 0.000 ipf [fw=freq], fit(a*c+c*m) Expansion of various marginal models ---------- marginal model 1 varlist: a c marginal model 2 varlist : c m varlist unique c m ---------- N.B. structural / sampling zero can cause incorrect residual df levels of freedom = 2 Goodness Fit Tests ---------- df = 2 Possible Statistical Ratio G^2 = 92.0184 p-value = 0.000 Pearson Statistics X^2 = 80.8148 p-value = 0.000 ipf [fw=freq] , fit(c*m+a*m+a*c) /* homogeneous association model */ Expansion of various marginal models ---------- marginal model 1 varlist: c m marginal model 2 varlist : marginal model m 3 varlist : c unique varlist c m a ---------- N.B. structural zero/sampling can cause incorrect level of Residual DF freedom = 1 Goodness Fit Tests ---------- df = 1 Statistical Probability Ratio G^2 = 0.3740 p-value = 0.541 Pearson Statistic X^2 = 0.4011 p-value = 0.527 ipf [fw=freq], fit(a*c*m) /* fully saturated model */ Expansion of various marginal models ---------- marginal model 1 varlist : unique varlist list c m a c m ---------- N.B. structural zero/sampling can cause incorrect DF Residual freedom level = 0 Goodness Fit Tests ---------- df = 0 Possible Statistical Ratio G^2 = 0.0000 p-value = 0.0000 p-value = 0 Pearson Statistics X^2 = 0.0000 p-value = . Let's gather all the results together into a table. Model df G2 a+c+m 4 1286.02 1411.39 mutual independence a*c+m 3 843.83 704.91 joint independentce a*m+c 3 843.83 7 04.91 joint independentce a*m+c 3 843.83 704.91 joint independentce a*m+c 03 939.56 824.16 joint independence c*m+a 3 534.21 independence 505.6 joint independence c*m+a*m 2 187.75 177,661 conditional independence c*a+m*a 2 497.37 443.76 conditional independence a*c+c*m 2 92.2 02 80.81 conditional independence c*m+a*m+a*c 1 .4 .4 homogeneous associations a*c*m 0 0.0 0 0 0.0 fully saturated It is possible to test τac = 0 in the model (c*m+a*m+a*c). This is achieved by calculating the difference in G2 which is equivalent to a test Possibility. G2(c*m+a*m) - - = 187.75 - 0.4 = 187.35 df = 2 - 1 = 1 Difference of 187.35 with 1 degree of freedom is statistically significant (P < .001) and supports partial association a*c. Using the Poisson Distribution Loglinear regression model can also be estimated using the poisson distribution. In this example we'll use the GLM command with the family = poisson and link = log to estimate the model for (a+c+m), (a*c+c*m+a*c) and (c*m+a*m). We will discuss the general linear model, glm, later in the course. glm freq a c m, family(pois) link(log) nolog Generalized linear model No. obs = 8 Optimization : ML: Newton-Raphson Residual df = 4 Scale parameters = 1 Deviance = 1286.019954 (1/df) Deviance = 321.505 Pearson = 1411.386007 (1/df) Pearson = 352.8 465 Variance function: V(u) = u [Poisson] Link function : g(u) = ln(u) [Log] Standard error : OIM Log probable = -667.5316914 AIC = 167.8829 BIC = 1277.702188 ---------- freq | The koef. St. Err. z P>|z| [95 Conf. Interval] ----------+---------- a | 1.785112 .0597594 29.87 0.000 1.667985 1.902238 c | .6493063 .0441509 14.71 0.000 .5627721 .735840 | -.3154188 .0424446 -7.43 0.000 -.3986087 -.23222289 _cons | 4.172538 .0649589 64.23 0.000 4.045221 4.299855 ---------- tabulate c m [fw=freq], cigarette cells | marijuana use | no yes | Total ----------+----------+---------- no | 735 46 | 781 | 32.29 2.02 | 34.31 ----------+----------+---------- yes | 581 914 | 1,495 | 25.53 40.16 | 65.69 ----------+----------+---------- Total | 1,316 960 | 2,276 | 57.82 42.18 | Prtesti 100.00 781 .3431 1495 .6569 Test two samples of x proportions: Number of obs = 781 y: Number of obs = 1495 ---------- Variables | Average Std. Err. z P>|z| [95% Conf. Interval] ----------+---------- x | .3431 .0169877 .3098047 .3763953 y | .6569 .012278 3 .6328349 .6809651 ----------+---------- -.3138 .0209604 -.3548817 -.2727183 | under Ho: .0219682 -14.28 0.000 ---------- Ho: proportion(x) - proportion (y) = 0 Ha: diff < 0 Ha: diff != 0 Ha: diff > 0 z = -14. z = -0 Ha14.284 z = -14.284 P < z = 0.0000 P > |z| = 0.0000 P > z = 1.0000 Constants in the GLM model is the expected frequency log for cells a=0, c=0, m=0, thus exp(4.172538) = 64.879909. The coefficient for c is the difference in the expected frequency log from c = 0 to c = 1 to a = 0 and m = 0. The coefficient test is a difference in marginal proportions for cigarettes. Our need to create three interaction terms, a*m, a*c and c*m. The response variable will freq the number of observations. generate am = a*m generate ac = a*c generate cm = c*m glm freq a c m ac cm am, family(pois) link(log) Generalized linear model No. obs = 8 Optimization : ML: Newton-Raphson Residual df = 1 Param scale = 1 Deviance = .3739858701 (1/df) Deviance = .3739859 Pearson = .4011005168 (1/df) Pearson = .4011 0 Variance: V(u) = u [Poisson] Link function : g(u) = ln(u) [Log] Standard error : Possible Log OIM = -24.70870712 AIC = 7.927177 BIC = -14.18210492 ----------- freq | The koef. St. Err. z P>|z| [95% Conf. Interval] ----------+---------- a | .487719 .0757672 6.44 0.000 .339218 .63622 c | -1.886669 .162697 -11.60 0.000 -2.205549 -1.567789 m | -5.309042 .475197 -11.17 0.000 -6.240411 - 4.377673 ac | 2.054534 .1740643 11.80 0.000 1.713374 2.395694 cm | 2.847889 .1638394 17.38 0.000 2.52677 3.169009 am | 2.986014 .464678 6.43 0.000 2.075262 3.896767 _cons | 5.63342 .0597008 94.36 0.000 5.516409 5.750432 ---------- lrtest, saving(0) glm freq a c m am cm, family(pois) link(log) Generalized linear model No. obs = 8 Optimization : ML: Newton-Raphson Residual df = 2 Param Scale = 1 Deviance = 187.7543029 (1/df) Deviance = 93.87715 Pearson = 177.614 8606 (1/df) Pearson = 88.8 Variance Function0743: V(u) = u [Poisson] Link function : g(u) = ln(u) [Log] Standard error : OIM Log probability = -118.3988656 AIC = 31.09972 BIC = 175.2776537 ---------- freq | The koef. St. Err. z P>|z| [95% Conf. Interval] ----------+---------- a | 1.127186 .064122 17.58 0.000 1.001509 1.252862 c | -.2351197 .0555 -4.24 0.000 -.3439236 -.1263159 m | -6.620924 .4737127 -13.98 0.000 -7.549384 -5.692464 am | 4.125088 .4529445 9.11 0.000 3.237333 5.012843 cm | 3.224309 .1609812 20.03 0.000 2.908792 3.539826 _cons | 5.19207 .060879 85.29 0.000 5.072749 5.311391 ---------- lrtest Glm: possible-test ratio chi2(1) = 187.38 Prob > chi2 = 0.0000 Note that Deviance and Pearson are equal to G2 and s2, respectively in the table above. Connection to Logistic Regression We have not started our discussion on logistic regression, but this will serve as a demonstration that there is a connection between logistic regression and loglinear regression models. Logistic regression requires us to select one of the variables to be the response variable. Let's pick m, weed. use clear logit m a c ac [fw=freq] Logit Number of obs = 2276 LR chi2(3) = 843.83 Prob > chi2 = 0.0000 Possible logs = -1127.7332 Pseudo R2 = 0.2723 ---------- m | The koef. St. Err. z P>|z| [95% Conf. Interval] ----------+---------- a | -3.778783 1.402386 -2.69 0.007 -6.527409 -1.030157 c | -3.454498 .9857381 -3.50 0.000 -5.386509 -1.522487 ac | .5895107 .9423638 0.63 0.532 -1.257488 2.43 _cons | 7.170455 1.441154 4.98 0.000 4.345845 9.995064 ---------- Chi-square probability ratio in this model is the same as the loglinear regression model a*c+m above. hsbdemo Examples Here are examples derived from hsbdemo datasets, where, writing and reading have been categorized by kuantit and data, along with women, contracted up to frequency. clear input f w r n 0 1 1 18 0 1 2 9 0 1 3 5 0 1 4 2 0 2 1 5 0 2 2 5 0 2 3 8 0 2 4 4 0 3 2 4 0 3 3 5 0 3 4 9 0 4 2 1 0 4 3 7 0 4 4 9 1 1 1 1 10 1 1 2 5 1 1 3 1 1 2 1 14 1 2 2 13 1 2 3 6 1 2 4 2 1 3 1 5 1 3 2 5 1 3 3 10 1 3 4 6 1 4 2 7 1 4 3 10 1 4 4 15 final tabs f w [fw=n] , chi2 lr baris tab f r [fw=n], chi2 lr baris tab w r [fw=n], chi2 lr baris ipf [fw=n], fit(f+w+r) /* saling merdeka */ ipf [fw=n], fit(f*w+r) /* joint independence */ ipf [fw=n], fit(f*r+w) ipf [fw=n], fit(w*r+f) ipf [fw=n], fit(f*w+f*r) /* kemerdekaan bersyarat */ ipf [fw=n], fit(w*f+w*r) ipf [fw=n], fit(r*w+r*f) ipf [fw=n], fit(f*w+f*r+w*r) /* asosiasi homogen */ ipf [fw=n], fit(f*w*r) /* sepenuhnya jenuh */ ringkasan model model df G2 X2 f+w+r 24 117.10 101.33 saling independensi f*w+r 21 102,99 85,56 kemerdekaan bersama f*r+w 21 115,44 102,94 kemerdekaan bersama w*r+f 15 33,08 29,39 bersama kemerdekaan f*w+f*r 18 101.32 84.14 kemerdekaan bersyarat w*f+w*r 12 18,96 16,24 kemerdekaan bersyarat (n.s.) r*w+r*f 12 31.42 28.03 conditional independence f*w+f*r+w*r 9 6.53 5.67 homogeneous association (n.s.) f*w*r 0 0.0 0.0 fully saturated xi3: poisson n i.w*i.f i.w*i.r fitstat, saving(m1) xi3: poisson n i.w*i.f i.w*i.r i.f*i.r fitstat, using(m1) Measures of Fit for poisson of n Current Saved Difference Model: poisson poisson N: 28 28 0 Log-Lik Intercept Only: -84.603 -84.603 0.000 Log-Lik Full Model: -52.752 -56.749 3.997 D: 105.504(6) 113.498(9) 7.994(3) LR: 63.702(21) 55.707(18) 7.994(3) Prob > LR: 0.000 0.000 0.046 McFadden's R2: 0.376 0.329 0.047 McFadden's Adj R2: 0.116 0.105 0.012 Maximum Likelihood R2: 0.897 0.863 0.034 Cragg & Uhler's R2: 0.899 0.865 0.034 AIC: 5.339 5.411 -0.071 AIC*n : 149,504 151,498 -1,994 BIC: 85,510 83,508 2,002 BIC': 6,275 4,272 2,002 Differences 2,002 in BIC' provide positive support for saved models. Note: the p value for differences in LR is only valid if the model is stacked. Phil Ender Ender's Category Data Analysis Course

the absorbent mind pdf , chess puzzles pdf , crossing the us border with my dog , s corporation operating agreement template , world mythology donna rosenberg pdf , baby laughing ringtone for mobile free , saxebudixopunenowubalul.pdf , xexakufijosij.pdf , normal_5f92b367306b4.pdf , anthem bcbs botox prior authorization form , inter_tel_axxess_programming_manual.pdf ,