On Pearson's X^2 for categorical response variables

Jeroen Weesie

Department of Sociology
University of Utrecht
The Netherlands

email J.Weesie@fss.uu.nl
tel +31–30–2531922
fax +31–30–2534405

Prepared for

Belgian/Dutch/German Stata User's Meeting
Maastricht May 23, 2002

Testing goodness of fit

- embedd model in larger model
  - constructive method: suggests model improvement

    but: often violation of some assumption leads to rejections for other forms of misspecification

    e.g.: White test for heteroscedasticity in regression is also very sensitive to misspecification of mean
  - may require estimating more complicated models (LR, Wald), though sometimes score testing may be feasible

    Check `testomit`
  - sometimes saturated model can be estimated, e.g., regression models with categorical covariates
- goodness of fit statistics
  - derive distribution of d(obs,fit) under Ho
  - example with categorical response

    d(obs,fit) = sum $(\text{obs-exp})^2/\text{exp}$
  - d() can often be seen as an aggregate of residuals
  - See Cressie-Read (1984) for details
- *All* models are wrong …

Pearson's X2 for binary data

Pearson X^2 measure for goodness of fit after binary regression models

$$\pi_i = F(x_i'\beta)$$

With **replication** of the x_i (HL: "m-asymptotics")

$$X^2 = \sum_{\text{pattern}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

With large number of obs per pattern, X^2 is approximately χ^2 (with df = #patterns - #parameters).

Stata's `lfit` command provides this test for logistic regression

`lfit` also allows essentially unique covariates, i.e., with small number of replications per pattern. The manual warns that this is "not necessarily incorrect."

Unaggregated measure of fit

With **unique covariates**, the unaggregated Pearson's statistic T_n is

$$T_n = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

With replicated data, T_n does not equal X^2, but usually is close.

Claim: T_n is *not* χ^2 distributed ("n-asymptotics").

Correct Theory: Subject to regularity conditions (Windmeijer '90; McCullagh '86)

$$\frac{T_n - n}{\sqrt{n}\,\sigma_n} \to N(0,1)$$

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-2\pi_i)^2}{\pi_i(1-\pi_i)} - v_n' \Omega_n^{-1} v_n$$

$$v_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i(1-\pi_i)} F'(x_i'\beta)x_i$$

$$\Omega_n = \frac{1}{n}\sum_{i=1}^{n} \frac{F'(x_i'\beta)}{\pi_i(1-\pi_i)} \qquad \text{Fisher information}$$

Condition for T_n to be χ^2 distributed: $\sigma^2 = 2$.

Counter example: logistic regression with 1 x-var

$$x_i \sim U[-1,2] \qquad \beta = 1$$
$$\sigma_n^2 \to 0.034$$

Extensions available for (Windmeijer 1995)

- *multinomial logit* (Stata: `mlogit`)
- *conditional logistic regression* (one success/group; = Luce-McFadden choice model) (Stata: `clogit`).

Conditional logistic regression (k alternatives)

$$\pi_{ij} = \frac{\exp x_{ij}\beta}{\Sigma_{h=1}^{k} \exp x_{ih}\beta}$$

Asymptotic result for

$$T_n = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(Y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}$$

$$\frac{T_n - nk}{\sqrt{n}\,\sigma_n} \to N(0,1)$$

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \left\{ \frac{1 - 2\pi_{ij}}{\pi_{ij}} q_{ij} - \sum_{h \neq j}^{k} q_{ij} q_{ih} \right\} - \nu_n' \Omega_n^{-1} \nu_n$$

$$\nu_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} q_{ij} (x_{ij} - \sum_{h=1}^{k} \pi_{ih} x_{ih})$$

$$q_{ij} = \frac{1 - 2\pi_{ij}}{1 - \pi_{ij}}$$

Problem with application

(1) Sensitivity with respect to observations with large residuals (small π_i for observed response)

Ad hoc modifications of test statistics

- ignore observations with some $\pi < \epsilon$
- "round-up" probabilities to ϵ
- or: "leave as is"

See also Hosmer & Lemeshow - 2nd edition.

(2) Quality of asymptotic approximation unknown

(3) Power against meaningfull misspecifications unknown

A Stata command

```
Post-estimation command

    pearsonx2 [, eps(#) table]

available after the following commands

    logit / logistic
    probit
    cloglog
    mlogit
    clogit -- one positive response per groups

Options

eps(#) specifies that only observations for which the
    estimated probability for all possible outcomes are
    greater than # are used in computing the test.
    # defaults to 1E-2.

table specifies that Windmeijer's test is conducted for
    various eps (.1,.01,.001,etc) in order to assess the
    sensitivity of the test to very small probabilities
    of some outcomes.
```

Example logistic regression

```
. use barcelona_lbw
(Hosmer & Lemeshow data)
. xi: logistic low age lwt i.race smoke ptl ht ui
i.race            _Irace_1-3         (naturally coded; _Irace_1 omitted)

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(8)      =      33.22
                                                  Prob > chi2     =     0.0001
Log likelihood =   -100.724                       Pseudo R2       =     0.1416

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9732636   .0354759    -0.74   0.457     .9061578    1.045339
         lwt |   .9849634   .0068217    -2.19   0.029     .9716834    .9984249
    _Irace_2 |   3.534767   1.860737     2.40   0.016     1.259736    9.918406
    _Irace_3 |   2.368079   1.039949     1.96   0.050     1.001356    5.600207
       smoke |   2.517698    1.00916     2.30   0.021     1.147676    5.523162
         ptl |   1.719161   .5952579     1.56   0.118     .8721455    3.388787
          ht |   6.249602   4.322408     2.65   0.008     1.611152    24.24199
          ui |     2.1351   .9808153     1.65   0.099     .8677528      5.2534
------------------------------------------------------------------------------

. lfit

Logistic model for low, goodness-of-fit test

       number of observations =       189
 number of covariate patterns =       182
            Pearson chi2(173) =       179.24
                  Prob > chi2 =         0.3567

. lfit, group(10)

Logistic model for low, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

       number of observations =       189
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         9.65
                  Prob > chi2 =         0.2904
```

Logistic regression (cont)

```
 . pearsonx2

Pearson-Windmeijer goodness-of-fit test after logistic low

       number of observations =      189
     Pearson's X2 (ungrouped) =   182.02
    Windmeijer's H = norm(X2) =     0.61
               Prob > chi2(1) =   0.4334

. pearsonx2, table

Pearson-Windmeijer goodness-of-fit test after logistic low

       number of observations =      189
     Pearson's X2 (ungrouped) =   182.02
    Windmeijer's H = norm(X2) =     0.61
               Prob > chi2(1) =   0.4334

----------------------------------------------------------------
       eps  |   Obs         X2      se(X2)         H          p
-----------+----------------------------------------------------
0.10000000 |   163     161.46        3.01       0.26    0.6095
0.01000000 |   189     182.02        8.90       0.61    0.4334
----------------------------------------------------------------
       All obs with some p<eps are ignored in computing the test
```

Example conditional logistic regression

```
clogit choice sexJap incJap japan sexEur incEur europe, group(id) nolog

Conditional (fixed-effects) logistic regression   Number of obs   =        885
                                                  LR chi2(6)      =     142.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -252.72012                       Pseudo R2       =     0.2202

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexJap |  -.4694799   .3114939    -1.51   0.132    -1.079997     .141037
      incJap |   .0276854   .0123666     2.24   0.025     .0034472    .0519236
       japan |  -1.962652   .6216804    -3.16   0.002    -3.181123   -.7441806
      sexEur |   .5388442   .4525278     1.19   0.234     -.348094    1.425782
      incEur |   .0273669    .013787     1.98   0.047      .000345    .0543889
      europe |  -3.180029   .7546837    -4.21   0.000    -4.659182   -1.700876
------------------------------------------------------------------------------

. pearsonx2, table

Pearson-Windmeijer goodness-of-fit test after clogit choice

       number of observations =      295
     Pearson's X2 (ungrouped) =   870.42
    Windmeijer's H = norm(X2) =    19.48
               Prob > chi2(1) =   0.0000

----------------------------------------------------------------
      eps   |  Obs        X2      se(X2)         H         p
-----------+----------------------------------------------------
0.10000000 |  200     625.43       2.26     126.67    0.0000
0.01000000 |  295     870.42       3.30      19.48    0.0000
----------------------------------------------------------------
       All obs with some p<eps are ignored in computing the test
```

Design

$$
\begin{aligned}
\text{True :} \quad & \text{logit}(\pi_i) = \gamma x_1 + \gamma x2 \\
& x_{ij} \text{ iid } N(0,1) \\
\text{Fitted:} \quad & \text{logit}(\pi_i) = \beta_1 x_1 + \beta_2 x2 \\
& \text{probit}(\pi_i) = \beta_1 x_1 + \beta_2 x2
\end{aligned}
$$

Results (proportion of rejections in 1000 replications)

			`pearsonx2`			`linktest`		
γ	n	fitted	0.10	0.05	0.01	0.10	0.05	0.01
1	100	logit	.048	.027	.015	.105	.049	.005
1	400	logit	.070	.045	.021	.100	.050	.008
1	1600	logit	.064	.046	.027	.119	.062	.013
1	100	probit	.046	.032	.018	.112	.055	.010
1	400	probit	.096	.076	.047	.120	.059	.009
1	1600	probit	.102	.080	.058	.147	.085	.021
3	100	logit	.105	.063	.028	.100	.064	.017
3	400	logit	.127	.071	.021	.142	.116	.058
3	1600	logit	.133	.074	.023	.103	.062	.014
3	100	probit	.160	.119	.055	.194	.112	.037
3	400	probit	.248	.178	.066	.253	.204	.134
3	1600	probit	.269	.201	.080	.254	.164	.078

Simulation: Omitted variables in logit

Design

$$\begin{array}{ll} \text{true} & \text{logit}\pi_i = x_{i1} + x_{i2} + \gamma x_{i1} x_{i2} \\ & x_{ij} \text{ iid } N(0,1) \\ \text{fitted} & \text{logit}\pi_i = \beta_1 x_{i1} + \beta_2 x_{i2} \end{array}$$

Results (proportion of rejections in 1000 replications)

n	γ	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
100	0	.037	.028	.020
400	0	.085	.053	.026
1600	0	.076	.037	.010
100	1/3	.087	.070	.053
400	1/3	.112	.089	.036
1600	1/3	.248	.181	.092
100	2/3	.230	.201	.161
400	2/3	.273	.194	.113
1600	2/3	.847	.781	.641
100	1	.488	.459	.394
400	1	.621	.536	.348
1600	1	.999	.999	.995

Simulation: Omitted variables in clogit

Design – k alternatives

$$\text{True} \quad \text{LP}_{ij} = x_{1ij} + x_{2ij} + \gamma x_{1ij} x_{2ij}$$
$$x_{hij} \text{ iid } N(0,1)$$

$$\text{Fitted} \quad \text{LP}_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij}$$

and

$$\pi_{ij} = \frac{\exp \text{LP}_{ij}}{\Sigma_{l=1}^{k} \exp \text{LP}_{il}}$$

Results (proportion of rejections in 1000 replications)

		$\gamma = 0$			$\gamma = 1$		
k	n	.100	.050	.010	.100	.050	.010
3	100	.031	.023	.017	.378	.352	.284
3	200	.040	.030	.017	.267	.211	.106
3	400	.046	.033	.024	.506	.475	.436
3	400	.057	.035	.021	.268	.194	.103
4	100	.061	.044	.026	.261	.229	.192
4	200	.022	.019	.006	.813	.747	.619
4	400	.051	.030	.017	.714	.636	.473
4	800	.059	.035	.019	.995	.920	.817
5	100	.052	.046	.024	.180	.133	.071
5	200	.028	.019	.007	.920	.905	.881
5	400	.061	.046	.022	.771	.722	.594
5	800	.046	.030	.014	.996	.994	.990

Discussion and conclusion

(Based on many more simulations than reported here)

Dedicated tests (eg omitted vars test) have more power than the omnibus gof test (surprise?)

Asymptotic results for binary cases (logit, probit) seem adequate

I am not sure yet about cloglog

Asymptotic results for mlogit / clogit are reasaonbly accurate only for LARGE n. For small and moderate n, tests are severely biased. Turn to higher order asymptotics?

The methods of Windmeijer (1994) and Weesie (199) for reducing the sensitivity of the tests to very small probabilities are not ambiguous improvements.

Consider other statistics from the power family suggested by Cressie-Read.

References

- Read, T.R.C. and N.A.C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New York: Springer Verlag.

  See also Stata command `multgof`, a revision of `gof` published in STB 36 1997.

- Hosmer, D.W. and S. Lemeshow. 1989. *Applied logistic regression.* 2nd edition. Wiley: New York.

- McCullagh, P. 1986. "The conditional distribution of goodness-of-fit statistics for discrete data." *Journal of the American Statistical Association* 81: 104–107.

- Weesie, J. 1998. "Windmeijer's goodness-of-fit test for logistic regression." Stata Technical Bulletin Reprints Vol 8, 153–160.

- Weesie, J. 1999 (2001). Tests for omitted variables. Utrecht, memeo.

  Describes the command `testomittedvars` that performs LR, Wald, and three forms of score tests for omitted variables.

- Windmeijer, F.A.G. 1990. "The asymptotic distribution of the sum of weighted squared residuals in binary choice models." *Statistica Neerlandica* 44, 2: 69–78.

- Windmeijer, F.A.G. 1994. "A goodness-of-fit test in the multinomial logit model based on weighted squared residuals." *Statistica Neerlandica* 48, 3: 271–283.