Assessing the fit of non-canonical binary regression models using the Hjort-Hosmer statistic

hh.ado

Steve Quinn,[1] David W Hosmer[2]

1. School of Medicine, Flinders University, Adelaide SA, Australia
2. Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst MA, USA

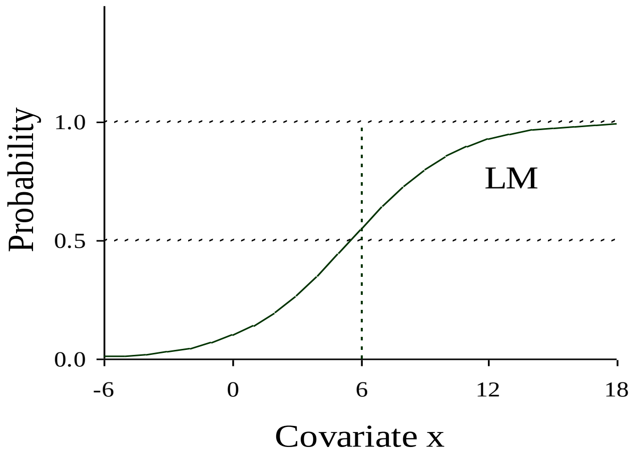
Flinders UNIVERSITY
inspiring achievement

Structure of the talk

- The setting - forms of binary regression
- Motivation - model validation
  - The Hosmer-Lemeshow statistic
  - The Hjort-Hosmer statistic
- Justification – Simulations
- Pseudo code
- Examples

Background – The logistic model

Logistic regression has long been the workhorse of statistical analysis of binary outcome (yes/no) data.



$$\Pr\left(Y_i = 1 \mid \mathbf{x}_i\right) = \pi(\mathbf{x}_i) = \frac{e^{x_i'\boldsymbol{\beta}}}{1 + e^{x_i'\boldsymbol{\beta}}}$$

- Outputs Odds Ratios ≈ RR
- Symmetric around y = 0.5

If $\boldsymbol{Z_i} = 1 - \boldsymbol{Y_i}$ then

$$\Pr(\boldsymbol{Y_i} = 1 \mid \boldsymbol{x_i}) = 1 - \Pr(\boldsymbol{Z_i} = 1 \mid \boldsymbol{x_i})$$

Flinders UNIVERSITY inspiring achievement

Hosmer-Lemeshow statistic

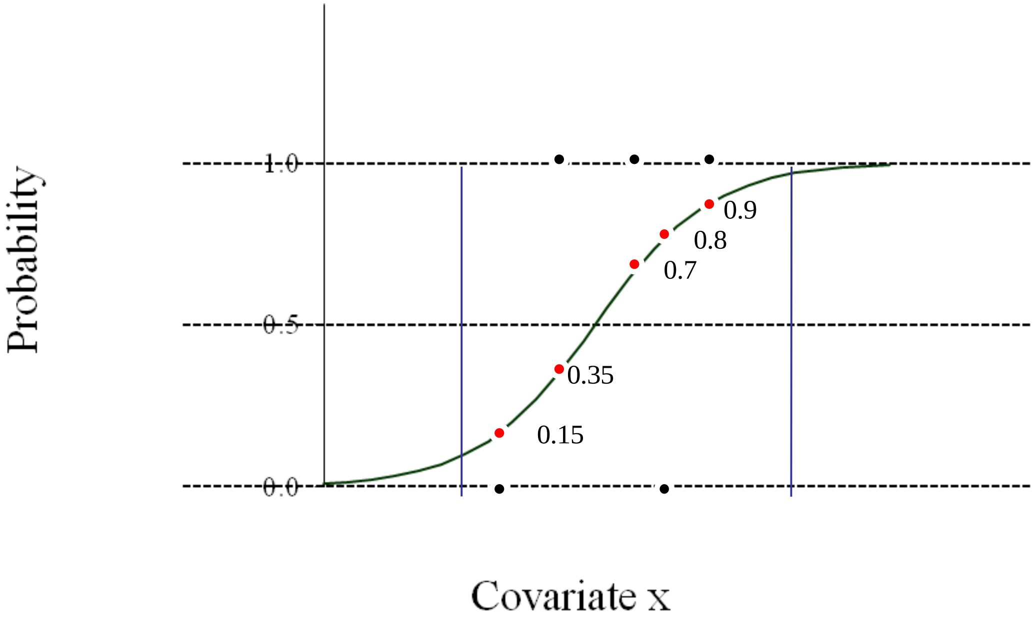
- Hosmer-Lemeshow "deciles-of-risk" test,

Hosmer, D. W. and S. Lemeshow (1980). "A goodness-of-fit test for the multiple logistic regression model." Communications in statistics **A10**: 1043-1069.

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)} \qquad \hat{C} \sim \chi^2_{g-2}$$

Normally, 10 groups

Hosmer-Lemeshow statistic



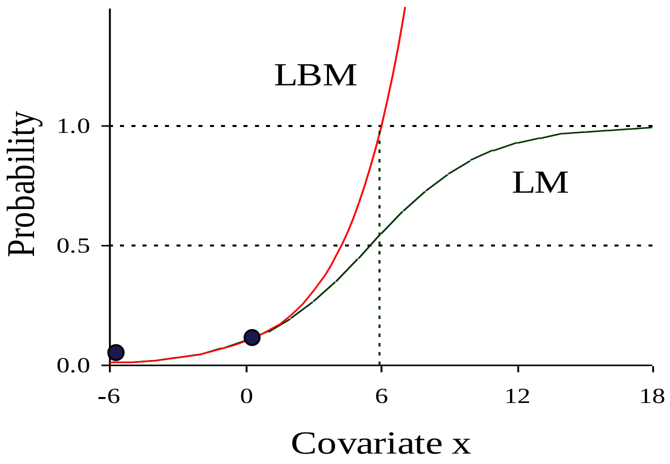
$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

$$\hat{C}_i = \frac{(3 - 5*0.5)^2}{5*0.5*(1-0.5)} = 0.2$$

Background – The log binomial model

Log link

$$\Pr\left(Y_i = 1 \mid \mathbf{x}_i\right) = \pi(\mathbf{x}_i) = e^{x_i'\boldsymbol{\beta}}$$



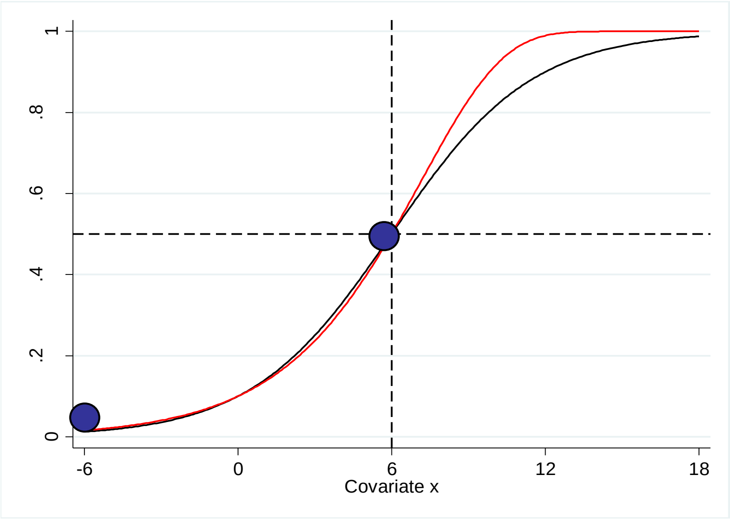
- Not symmetric
- Estimation algorithm can fail to converge
- Can produce inadmissible solutions

- Outputs RR

Complementary log-log model

Complementary log-log link

$$\Pr\left(Y_i = 1 \mid \mathbf{x}_i\right) = \pi(\mathbf{x}_i) = 1 - e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$



- Complementary log-log link
- Not symmetric
- Coefficients not interpretable.

Why bother?

- It has been used to calculate prevalence ratios (vs. prevalence odds ratios)

Bhattacharya R, Shen C, Sambamoorthi U, *Excess risk of chronic physical conditions associated with depression and anxiety*. BMC psychiatry. 14(2014), pp. 10.

- It has been used based on a biological expectation of an asymmetrical relationship between the systematic and random components

Gyimah SO, Adjei JK, Takyi BK, Religion, contraception, and method choice of married women in Ghana. Journal of religion and health. 51(4) (2012), pp. 1359-1374.

Why these 2 statistics?

- Goodness-of-fit (GOF) measures have been studied extensively for logistic regression models (currently Hosmer-Lemeshow used)

- GOF for the log binomial regression published in 2013 (recommended using Hjort-Hosmer)

Quinn SJ, Hosmer DW, Blizzard L, Goodness-of-fit statistics for log-link regression models. J Stat Comp Sim. 85(12) (2014), pp. 2533-2545

- CLL models currently being studied (similarly looks like recommending Hjort-Hosmer)

Hjort-Hosmer statistic

Hjort-Hosmer statistic

Hosmer DW, Hjort NL, (2002). “Goodness-of-fit processes for logistic regression: simulation results.” Statistics in medicine. 21(18), 2723-2738.

Based on partial sums of residuals, sorted by their fitted values.

Absolute maximal partial sum |M| are calculated.

Rationale: If the model is well-fit, then |M| is small.

What is a small |M|?

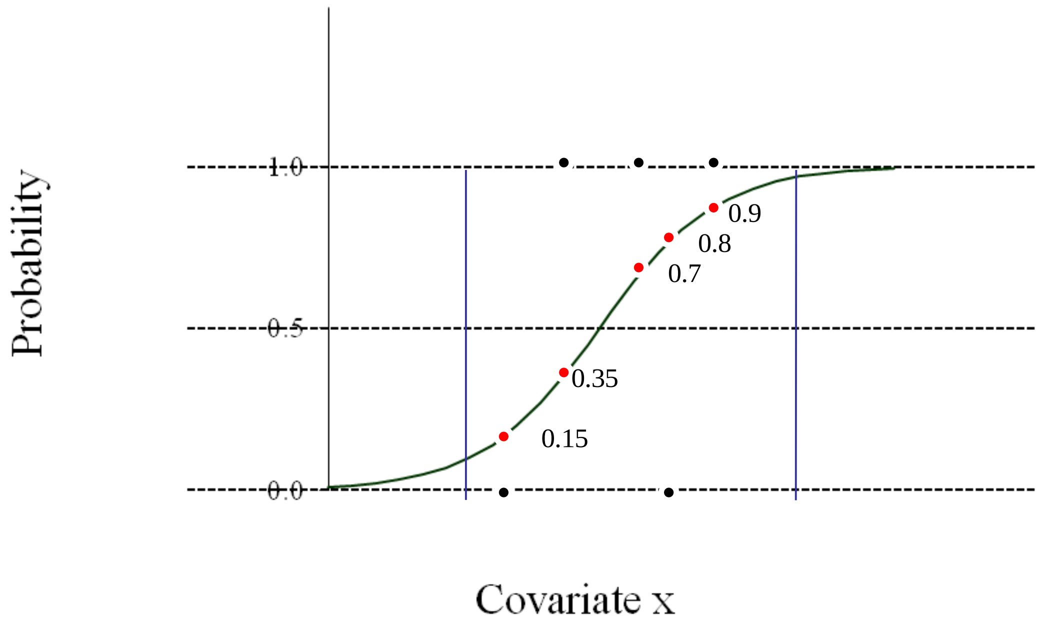
| M | is compared to n secondary partial sums $|M_j|$, each from a "correct" model:

a) comprises the same vector of covariates

b) outcomes simulated using that vector of covariates.

$$\text{P-value} = \sum_{\mathbf{j}} \mathbf{I_j}(|M_j| - |M|)/\mathbf{n}.$$

Hjort-Hosmer statistic



Residuals	Partial sums
–0.15	–0.15
0.65	0.50
0.30	0.80
–0.85	–0.05
0.10	0.05

$|M| = 0.8$

Performance of the statistics

1. Simulation the outcomes

   a) Start with a vector of covariates $\boldsymbol{x} \in \boldsymbol{U}(0,10),\ \boldsymbol{d} = 0,1$

   b) Take 200 random draws of $\boldsymbol{x}, \boldsymbol{d}$

   c) Specify the mean of a distribution function

$$\Pr\left(Y_i = 1 \mid \mathbf{x}_i, \beta_0, \beta_1, \beta_2\right) = \pi(\mathbf{x}_i) = 1 - e^{-e^{\beta_0 + \mathrm{x}_i'\beta_1 + d_i'\beta_2}}$$

Performance of the statistics

d) Predict outcomes

$$Y_i = \begin{cases} 1 \text{ if } 1 - e^{-e^{\beta_0 + \mathrm{x}_i'\beta_1 + d_i'\beta_2}} > u \\ 0 \text{ if } 1 - e^{-e^{\beta_0 + \mathrm{x}_i'\beta_1 + d_i'\beta_2}} < u \end{cases} \qquad \text{for } \boldsymbol{u} \in \boldsymbol{U}(0,1)$$

Performance of the statistics

2) Regress on either the correct or incorrect vector of covariates

```
cloglog y x d
```
correct - testing under the null

```
cloglog y x
```
incorrect - testing power

Apply each statistic to the regression

3) Repeat steps (1) and (2) 1000 times and count the number of rejections by each statistic

Three scenarios considered

1. The correct model – CLL regress $\boldsymbol{Y}$ on $\boldsymbol{x},\boldsymbol{d}$

2. Power (by omitting terms) – CLL regress $\boldsymbol{Y}$ on $\boldsymbol{x}$

3. Power (wrong link)

determine outcomes by $$Y_i = \begin{cases} 1 \text{ if } \dfrac{e^{\beta_0 + \mathrm{x}_\mathrm{i}'\beta_1 + d_i'\beta_2}}{1+e^{\beta_0 + \mathrm{x}_\mathrm{i}'\beta_1 + d_i'\beta_2}} > u \\ 0 \text{ if } \dfrac{e^{\beta_0 + \mathrm{x}_\mathrm{i}'\beta_1 + d_i'\beta_2}}{1+e^{\beta_0 + \mathrm{x}_\mathrm{i}'\beta_1 + d_i'\beta_2}} < u \end{cases}$$

CLL regress $\boldsymbol{Y}$ on $\boldsymbol{x},\boldsymbol{d}$

Power under the null – the correct model

Table 1. Per cent rejection - CLL
N= 200 or 600 with 1000 replications

1 continuous covariate		Goodness-of-fit statistics‡	
P(Y=1|x=10)*	Distribution	HL	HH
0.9	*U(0,10)*	7.4	5.5
0.1	*U(0,10)*	1.2	2.2
0.999	*N(5,3)*	6.4	6.4
0.5	$\chi(1)$	1.9	0.4
0.9	*U(0,10)*	6.8	5.1
0.1	*U(0,10)*	3.2	3.7
0.999	*N(5,3)*	7.2	5.3
0.5	$\chi(1)$	8.1	5.8
		5.3	4.3

*The curve also passes through P(Y=1|x=0) = 0.001

Power under the null – the correct model

Table 2. Per cent rejection - CLL
N= 200 or 600 with 1000 replications

1 continuous covariate + 1 dichotomous			Goodness-of-fit statistics‡	
P(Y=1|x=10,d=0)	P(Y=1|x=10,d=1)	Distribution	HL	HH
0.999	0.5	*U(0,10)*	6.6	5.0
0.999	0.5	*N(5,3)*	9.0	5.5
0.5	0.25	$\chi(1)$	2.7	6.1
0.5	0.25	$\chi(5)$	1.0	4.6
0.999	0.5	*U(0,10)*	8.0	5.4
0.999	0.5	*N(5,3)*	5.8	5.7
0.5	0.25	$\chi(1)$	7.7	5.5
0.5	0.25	$\chi(5)$	7.9	3.7
			6.1	5.2

*The curve also passes through P(Y=1|x=0,d=0) = 0.001

Flinders UNIVERSITY inspiring achievement

Power under the null – the correct model

Table 1. Per cent rejection - LB
N= 200 or 600 with 1000 replications

1 continuous covariate		Goodness-of-fit statistics‡	
P(Y=1|x=0)*	Distribution	HL	HH
0.1	*U(-6,6)*	5.1	4.5
0.3	*U(-6,2.1)*	4.8	5.2
0.5	*U(-6,1)*	3.9	4.2
0.7	*U(-6, 0.5)*	4.8	5.4
0.1	*U(-6,4)*	3.9	4.7
0.3	*U(-6,1)*	4.2	4.8
0.5	*U(-6,0)*	4.3	6.0
0.7	*U(-6, -0.5)*	4.6	4.5
		4.3	5.0

*The curve also passes through P(Y=1|x=-6) = 0.01

Power under the null – the correct model

Table 2. Per cent rejection - LB
N= 200 or 600 with 1000 replications

1 continuous covariate + 1 dichotomous		Goodness-of-fit statistics‡	
P(Y=1|x,d=0)	Distribution	HL	HH
0.1	*U(-6,4.18)*	*4.2*	*5.2*
0.3	*U(-6,0.9)*	*5.2*	*5.3*
0.5	*U(-6,0)*	*3.7*	*4.7*
0.7	*U(-6,-0.5)*	*4.0*	*4.5*
0.1	*U(-6,2.4)*	*3.7*	*3.4*
0.3	*U(-6,-0.25)*	*3.4*	*5.3*
0.5	*U(-6,-1)*	*2.9*	*4.1*
0.7	*U(-6,-1.5)*	*4.7*	*5.1*
		4.0	*4.7*

*The curve also passes through P(Y=1|x=-6,d=0) = 0.01 and P(Y=1|x=-6,d=1) = 0.02

Power under the alternative – an incorrect model

Table 3. Power – CLL
N=200 or 600 with 1000 replications

1 continuous + 1 continuous² covariate			Goodness-of-fit statistics‡	
P(Y=1|x=5)	P(Y=1|x=10)	Distribution	HL	HH
0.5	0.999	*U(0,10)*	15.2	17.1
0.3	0.5	*U(0,10)*	57.2	85.3
0.75	0.999	*N(5,3)*	13.1	15.3
0.75	0.999	$\chi(1)$	6.3	13.4
0.5	0.999	*U(0,10)*	38.7	40.5
0.3	0.5	*U(0,10)*	99.1	100
0.75	0.999	*N(5,3)*	5.0	35.3
0.75	0.999	$\chi(1)$	15.5	29.9
			31.3	42.1

*The curve also passes through P(Y=1|x=0, x^2=0) = 0.001

Power under the alternative – an incorrect model

Table 3. Power – LB
N=200 or 600 with 1000 replications

1 continuous + 1 continuous^2 covariate		Goodness-of-fit statistics‡	
P(Y=1|x=-6)	Distribution	HL	HH
0.01	*U(-6, 1.5)*	3.6	5.6
0.1	*U(-6, 1.5)*	6.8	10.6
0.3	*U(-6,1.5)*	19.7	32.9
0.5	*U(-6, 1.5)*	43.5	57.4
0.01	*U(-6, 3.11)*	3.2	6.9
0.1	*U(-6, 3.11)*	20.2	22.7
0.3	*U(-6, 3.11)*	76.6	83.6
0.5	*U(-6, 3)*	96.7	98
		33.8	39.7

*The curve also passes through P(Y=1|x=1.5) = 0.5 and P(Y=1|x=3) = 0.95

Power under the alternative – an incorrect model

Table 4. Power – CLL
N=200 or 600 with 1000 replications

1 continuous + 1 dichotomous covariate		**+ interaction**	Goodness-of-fit statistics‡	
P(Y=1|x=10,d=0)	P(Y=1|x=10,d=1)	Distribution	HL	HH
0.999	0.25	*U(0,10)*	19.3	5.9
0.999	0.5	*N(5,3)*	12.1	33.2
0.999	0.5	$\chi(3)$	13.2	6
0.5	0.25	$\chi(5)$	3.8	21.1
0.999	0.25	*U(0,10)*	28.5	12.9
0.999	0.5	*N(5,3)*	52.7	83.1
0.999	0.5	$\chi(3)$	22.4	5.1
0.5	0.25	$\chi(5)$	8.9	17.2
			5.5	10.9

*The curve also passes through P(Y=1|x=0,d=0) = 0.001

Power under the alternative – an incorrect model

Table 4. Power – LB
N=200 or 600 with 1000 replications

1 continuous + 1 dichotomous + interaction covariate		Goodness-of-fit statistics‡	
P(Y=1|x=3,d=1)	Distribution	HL	HH
0.3	*U(-6,3)*	4.1	4.1
0.5	*U(-6,3)*	4.0	4.5
0.7	*U(-6,3)*	2.9	4.2
0.9	*U(-6,3)*	4.3	5.3
0.3	*U(-6, 12.8)*	4.0	4.6
0.5	*U(-6, 6.8)*	3.7	5.0
0.7	*U(-6, 4.6)*	4.7	5.6
0.9	*U(-6, 3.4)*	4.2	5.3
		4.0	4.8

*The curve also passes through P(Y=1|x=-6,d=0) = 0.1 and P(Y=1|x=-6,d=1) = 0.1 and P(Y=1|x=0,d=0) = 0.2

Power under the alternative – an incorrect link

Table 5. Power – CLL
N=200 or 600 with 1000 replications

1 continuous covariate		Goodness-of-fit statistics‡	
P(Y=1|x=10,d=0)	Distribution	HL	HH
0.999	*U(0,10)*	22.7	27.1
0.9	*U(0,10)*	1.6	8.9
0.999	*N(5,5)*	29.9#	21.7
0.999	$\chi(1)$	5.0	5.4
0.999	*U(0,10)*	61.4	69
0.9	*U(0,10)*	6.2	20
0.999	*N(5,5)*	94.5#	50.2
0.999	$\chi(1)$	4.3	11.6
# only 87 (n=200) and 37 (n=600) of 1000 replications produced a test statistic		16.9	26.7

*The curve also passes through P(Y=1|x=0,d=0) = 0.001

Flinders UNIVERSITY inspiring achievement

Power under the alternative – an incorrect link

Table 5. Power – LB
N=200 or 600 with 1000 replications

1 continuous covariate		Goodness-of-fit statistics‡	
Link	Distribution	HL	HH
cloglog	*U(-6,6)*	3	5.3
logit	*U(-6,6)*	2.6	6.5
probit	*U(-6,6)*	0.7	12.8
cloglog	*U(-6,8)*	7.1	26.3
logit	*U(-6,9)*	7.9	28.2
probit	*U(-6,9)*	35.2	85.5
		9.4	27.4

*The curve passes through P(Y=1|x=-6) = 0.01 and P(Y=1|x=1.5) = 0.1

Power under the alternative – an incorrect link

Table 6. Power – CLL
N=200 or 600 with 1000 replications

1 continuous + 1 dichotomous covariate			Goodness-of-fit statistics‡	
P(Y=1|x=10,d=0)	P(Y=1|x=10,d=0)	Distribution	HL	HH
0.999	*0.5*	*U(0,10)*	7.1	13.3
0.9	*0.5*	*U(0,10)*	2.4	6.6
0.999	*0.5*	*N(5,5)*	3.3	46.6
0.999	*0.5*	*N(5,1)*	7.5	12.5
0.999	*0.5*	*U(0,10)*	4.7	70.4
0.9	*0.5*	*U(0,10)*	2.7	13.3
0.999	*0.5*	*N(5,5)*	3.1	31.5
0.999	*0.5*	*N(5,1)*	21.8	37.1
			6.6	28.9

*The curve also passes through P(Y=1|x=0,d=0) = 0.001

Flinders UNIVERSITY
inspiring achievement

Power under the alternative – an incorrect link

Table 6. Power – LB
N=200 or 600 with 1000 replications

1 continuous + 1 dichotomous covariate		Goodness-of-fit statistics‡	
Link	Distribution	HL	HH
cloglog	*U(-6,4)*	3.4	5.6
logit	*U(-6,4)*	1.9	6.1
probit	*U(-6,4)*	1.4	9
cloglog	*U(-6,6)*	4.1	15.6
logit	*U(-6,6)*	4	12.1
probit	*U(-6,6)*	2.6	34.2
		2.9	13.8

*The curve passes through P(Y=1|x=-6,d=0) = 0.01 and P(Y=1|x=-6,d=1) = 0.02 and P(Y=1|x=0,d=10) = 0.1

Positives of each statistic

Hosmer-Lemeshow (HL)

1. Easy to understand
2. In all major software programs today
3. Quick
4. Link invariant

Hjort-Hosmer (HH)

1. More precise
2. More powerful
3. Always produces a test statistic
4. Link invariant

Hjort-Hosmer statistic – pseudo code

1. Get the regression parameters
2. Calculate the absolute maximal partial sum |M|
3. Simulate outcomes based on the model covariates and the link function
4. Calculate a secondary maximal partial sum $|M_j|$
5. Repeat steps 3 and 4 “100” times.
6. Calculate a p-value $= \sum_{j} \mathbf{I_j}(|M_j| - |M|)/\mathbf{n}.$

The code

1. Implemented as an ado file, called hh.ado

2. Takes one argument – the number of repeated simulations

Example 1

```
. logistic foreign price

Logistic regression                               Number of obs   =         74
                                                  LR chi2(1)      =       0.17
                                                  Prob > chi2     =     0.6784
Log likelihood =  -44.94724                       Pseudo R2       =     0.0019
```

foreign	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
price	1.000035	.0000844	0.42	0.676	.9998699	1.000201
_cons	.339666	.1996674	-1.84	0.066	.1073214	1.075023

```
. hh 100
Hjort-Hosmer goodness-of-fit p-value = .25
```

Example 1

```
Logistic model for foreign, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =        74
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         5.39
                  Prob > chi2 =         0.7149
```

Example 2

```
. binreg foreign price, rr nolog

Generalized linear models                          No. of obs      =         74
Optimization     : MQL Fisher scoring              Residual df     =         72
                   (IRLS EIM)                      Scale parameter =          1
Deviance         =  89.91755851                    (1/df) Deviance =   1.248855
Pearson          =  73.93100323                    (1/df) Pearson  =   1.026819

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u)                    [Log]

                                                   BIC             =  -219.9751
```

foreign	Risk Ratio	EIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
price	1.000021	.0000567	0.37	0.712	.9999098	1.000132
_cons	.260734	.1060397	-3.31	0.001	.1174943	.5786004

```
. hh 100
Hjort-Hosmer goodness-of-fit p-value = .18
```

Questions or comments ?

Flinders
UNIVERSITY

inspiring achievement
www.flinders.edu.au