

15) Poisson Negative Binomial Regression

Vitor Kamada

August 2018

Poisson Distribution

$$Pr(Y = y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, \dots$$

$$E(Y) = Var(Y) = \mu$$

quietly set obs 10000

set seed 7

generate $Y = \text{rpoisson}(1)$

summarize Y

tabulate Y

$Y \sim \text{Poisson}(1)$

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	10,000	1.0038	1.000143	0	7

Y	Freq.	Percent	Cum.
0	3,660	36.60	36.60
1	3,679	36.79	73.39
2	1,861	18.61	92.00
3	603	6.03	98.03
4	164	1.64	99.67
5	27	0.27	99.94
6	5	0.05	99.99
7	1	0.01	100.00
Total	10,000	100.00	

$$Pr(Y = 0 | \mu = 1) = e^{-1} = 0.368$$

$$Pr(Y = y) = \frac{e^{-\mu} \mu^y}{y!}$$

$$E[y|x] = \mu_i = \exp(x_i' \beta)$$

$$\ln L(\beta) = \sum_{i=1}^N \{y_i x_i' \beta - \exp(x_i' \beta) - \ln y_i!\}$$

$$\mathbf{FOC:} \sum_{i=1}^N (y_i - \exp(x_i' \beta)) x_i = 0$$

$$V_{PML}[\hat{\beta}_p]$$
$$= \left(\sum_{i=1}^N \mu_i x_i x_i' \right)^{-1} \left(\sum_{i=1}^N w_i x_i x_i' \right) \left(\sum_{i=1}^N \mu_i x_i x_i' \right)^{-1}$$

$$w_i = V[y_i | x_i]$$

$$V[\hat{\beta}_p] = \left(\sum_{i=1}^N \mu_i x_i x_i' \right)^{-1}$$

Interpretation of Regression Coefficients

$$E[y|x] = \exp(x'\beta)$$

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j \exp(x'\beta)$$

$$AME = \hat{\beta}_j \frac{1}{N} \sum_{i=1}^N \exp(x_i' \hat{\beta})$$

If intercept is included, then $\hat{\beta}_j \bar{y}$

RAND Health Insurance Experiment (1974 to 1982)

Variable	Definition	Mean	Std. Dev.
MDU	Number of outpatient visits to an MD	2.861	4.505
LC	$\ln(\text{coinsurance} + 1)$, $0 \leq \text{coinsurance} \leq 100$	1.710	1.962
IDP	1 if individual deductible plan, 0 otherwise	0.220	0.414
LPI	$\ln(\max(1, \text{annual participation incentive payment}))$	4.709	2.697
FMDE	0 if $IDP = 1$ $\ln(\max(1, MDE/(0.01 \text{ coinsurance})))$ otherwise	3.153	3.641
LINC	$\ln(\text{family income})$	8.708	1.228
LFAM	$\ln(\text{family size})$	1.248	0.539
AGE	Age in years	25.718	16.768
FEMALE	1 if person is female	0.517	0.500
CHILD	1 if age is less than 18	0.402	0.490
FEMCHILD	$FEMALE * CHILD$	0.194	0.395
BLACK	1 if race of household head is black	0.182	0.383
EDUCDEC	Education of the household head in years	11.967	2.806
PHYSLIM	1 if the person has a physical limitation	0.124	0.322
NDISEASE	Number of chronic diseases	11.244	6.742
HLTHG	1 if self-rated health is good	0.362	0.481
HLTHF	1 if self-rated health is fair	0.077	0.267
HLTHP	1 if self-rated health is poor	0.015	0.121
	Omitted category is excellent self-rated health		

Cameron & Trivedi (2005)

sum MDU \$XLIST

Variable	Obs	Mean	Std. Dev.	Min	Max
MDU	20,186	2.860696	4.504765	0	77
LC	20,186	2.383588	2.041713	0	4.564348
IDP	20,186	.2599822	.4386354	0	1
LPI	20,186	4.708827	2.697293	0	7.163699
FMDE	20,186	4.030322	3.471234	0	8.294049
PHYSLIM	20,186	.1235247	.3220437	0	1
NDISEASE	20,186	11.2445	6.741647	0	58.6
HLTHG	20,186	.3620826	.4806144	0	1
HLTHF	20,186	.0772813	.2670439	0	1
HLTHP	20,186	.0149609	.1213992	0	1
LINC	20,186	8.708167	1.22841	0	10.28324
LFAM	20,186	1.248404	.5390681	0	2.639057
EDUCDEC	20,186	11.96681	2.806255	0	25
AGE	20,186	25.71844	16.76759	0	64.27515
FEMALE	20,186	.5169424	.4997252	0	1
CHILD	20,186	.4014168	.4901972	0	1
FEMCHILD	20,186	.1937481	.3952436	0	1
BLACK	20,186	.1815343	.3827365	0	1

poisson MDU \$XLIST

Poisson regression

Number of obs = 20,186

LR chi2(17) = 13106.07

Prob > chi2 = 0.0000

Pseudo R2 = 0.0983

Log likelihood = -60087.622

MDU	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
LC	-.0427332	.0060785	-7.03	0.000	-.0546469	-.0308195
IDP	-.1613169	.0116218	-13.88	0.000	-.1840952	-.1385385
LPI	.0128511	.0018362	7.00	0.000	.0092523	.0164499
FMDE	-.020613	.0035521	-5.80	0.000	-.027575	-.0136511
PHYSLIM	.2684048	.0123624	21.71	0.000	.2441749	.2926347
NDISEASE	.023183	.0006081	38.12	0.000	.0219912	.0243749
HLTHG	.0394004	.0095884	4.11	0.000	.0206074	.0581934
HLTHF	.2531119	.016212	15.61	0.000	.2213369	.2848869
HLTHP	.5216034	.0272382	19.15	0.000	.4682176	.5749892
LINC	.0834099	.0051656	16.15	0.000	.0732854	.0935343
LFAM	-.1296626	.0089603	-14.47	0.000	-.1472245	-.1121008
EDUCDEC	.0176149	.0016387	10.75	0.000	.0144031	.0208268
AGE	.0023756	.0004311	5.51	0.000	.0015306	.0032206
FEMALE	.3487667	.0113504	30.73	0.000	.3265203	.371013
CHILD	.3361904	.0178194	18.87	0.000	.3012649	.3711158
FEMCHILD	-.3625218	.0179396	-20.21	0.000	-.3976827	-.3273608
BLACK	-.6800518	.0155484	-43.74	0.000	-.7105262	-.6495775
_cons	-.1898766	.0491731	-3.86	0.000	-.2862541	-.093499

poisson MDU \$XLIST, robust

Poisson regression

Log pseudolikelihood = -60087.622

Number of obs = 20,186
 Wald chi2(17) = 1924.78
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.0983

MDU	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
LC	-.0427332	.0150712	-2.84	0.005	-.0722723	-.0131942
IDP	-.1613169	.0279441	-5.77	0.000	-.2160863	-.1065474
LPI	.0128511	.0044136	2.91	0.004	.0042007	.0215015
FMDE	-.020613	.0088874	-2.32	0.020	-.0380319	-.0031941
PHYSLIM	.2684048	.0325743	8.24	0.000	.2045604	.3322493
NDISEASE	.023183	.0017189	13.49	0.000	.019814	.0265521
HLTHG	.0394004	.023194	1.70	0.089	-.006059	.0848598
HLTHF	.2531119	.0429454	5.89	0.000	.1689405	.3372833
HLTHP	.5216034	.0748808	6.97	0.000	.3748398	.668367
LINC	.0834099	.0139182	5.99	0.000	.0561306	.1106891
LFAM	-.1296626	.0226793	-5.72	0.000	-.1741132	-.085212
EDUCDEC	.0176149	.004042	4.36	0.000	.0096927	.0255371
AGE	.0023756	.0011184	2.12	0.034	.0001837	.0045675
FEMALE	.3487667	.0283549	12.30	0.000	.293192	.4043413
CHILD	.3361904	.040411	8.32	0.000	.2569863	.4153945
FEMCHILD	-.3625218	.04415	-8.21	0.000	-.4490542	-.2759893
BLACK	-.6800518	.0368748	-18.44	0.000	-.7523252	-.6077785
_cons	-.1898766	.127516	-1.49	0.136	-.4398033	.0600502

margins, dydx(*)

Average marginal effects
Model VCE : **Robust**

Number of obs = 20,186

Expression : Predicted number of events, predict()

dy/dx w.r.t. : LC IDP LPI FMDE PHYSLIM NDISEASE HLTHG HLTHF HLTHP LINC LFAM EDUC
FEMCHILD BLACK

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
LC	-.1222467	.0431486	-2.83	0.005	-.2068164	-.037677
IDP	-.4614785	.08007	-5.76	0.000	-.6184128	-.3045442
LPI	.0367631	.0126336	2.91	0.004	.0120017	.0615244
FMDE	-.0589676	.0254221	-2.32	0.020	-.1087939	-.0091413
PHYSLIM	.7678244	.0944299	8.13	0.000	.5827453	.9529036
NDISEASE	.0663196	.0050145	13.23	0.000	.0564913	.0761478
HLTHG	.1127125	.0663211	1.70	0.089	-.0172744	.2426994
HLTHF	.7240761	.1232422	5.88	0.000	.4825259	.9656263
HLTHP	1.492149	.2146029	6.95	0.000	1.071535	1.912763
LINC	.2386102	.0399148	5.98	0.000	.1603787	.3168418
LFAM	-.3709253	.0651288	-5.70	0.000	-.4985754	-.2432753
EDUCDEC	.0503909	.0115909	4.35	0.000	.0276732	.0731087
AGE	.0067958	.0032025	2.12	0.034	.0005192	.0130725
FEMALE	.9977152	.0820078	12.17	0.000	.8369829	1.158448
CHILD	.9617383	.1167579	8.24	0.000	.732897	1.19058
FEMCHILD	-1.037064	.1272852	-8.15	0.000	-1.286539	-.78759
BLACK	-1.945421	.1081044	-18.00	0.000	-2.157302	-1.73354

Negative Binomial Distribution

$$\begin{aligned} &Pr(Y = y|\mu, \alpha) \\ &= \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1})\Gamma(y+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^y \end{aligned}$$

$$E(Y|\mu, \alpha) = \mu$$

$$Var(Y|\mu, \alpha) = \mu(1 + \alpha\mu)$$

generate $G = \text{rgamma}(1,1)$

generate $Z = \text{rpoisson}(G)$

summarize Z

tabulate Z

$$Z \sim \text{NB}(\mu = 1, \sigma^2 = 2)$$

Variable	Obs	Mean	Std. Dev.	Min	Max
z	20,186	.9896463	1.404738	0	13

z	Freq.	Percent	Cum.
0	10,149	50.28	50.28
1	5,040	24.97	75.25
2	2,502	12.39	87.64
3	1,302	6.45	94.09
4	582	2.88	96.97
5	287	1.42	98.39
6	151	0.75	99.14
7	93	0.46	99.60
8	44	0.22	99.82
9	17	0.08	99.91
10	11	0.05	99.96
11	5	0.02	99.99
12	2	0.01	100.00
13	1	0.00	100.00
Total	20,186	100.00	

Negative Binomial MLE

$$E[y|x] = \mu_i = \exp(x_i' \beta)$$

$$\sum_{i=1}^N \left(\frac{y_i - \mu_i}{1 + \alpha \mu_i} \right) x_i = 0$$

$$\sum_{i=1}^N \left[\frac{1}{\alpha^2} \left\{ \ln(1 + \alpha \mu_i) - \sum_{j=0}^{y_i-1} \frac{1}{(j + \alpha^{-1})} \right\} + \frac{y_i - \mu_i}{\alpha(1 + \alpha \mu_i)} \right] = 0$$

$$V[y_i|x_i] = \mu_i + \alpha g(\mu_i)$$

NB1 if $g(\mu) = \mu$

NB2 if $g(\mu) = \mu^2$

$H_0 : \alpha = 0$ (**equidispersion**)

nbreg MDU \$XLIST, robust

Negative binomial regression		Number of obs	=	20,186
		Wald chi2(17)	=	2203.12
Dispersion	= mean	Prob > chi2	=	0.0000
Log pseudolikelihood	= -42777.611	Pseudo R2	=	0.0320

MDU	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
LC	-.0504405	.0156238	-3.23	0.001	-.0810625	-.0198184
IDP	-.1475976	.0303777	-4.86	0.000	-.2071367	-.0880585
LPI	.0158351	.004431	3.57	0.000	.0071505	.0245197
FMDE	-.021335	.0090748	-2.35	0.019	-.0391211	-.0035488
PHYSLIM	.2751715	.0341067	8.07	0.000	.2083235	.3420195
NDISEASE	.0259352	.0016925	15.32	0.000	.022618	.0292524
HLTHG	.0065371	.023814	0.27	0.784	-.0401375	.0532118
HLTHF	.2368643	.0436579	5.43	0.000	.1512963	.3224322
HLTHP	.4256563	.0686042	6.20	0.000	.2911945	.560118
LINC	.0845165	.0113918	7.42	0.000	.0621891	.106844
LFAM	-.1226764	.0231639	-5.30	0.000	-.1680769	-.0772759
EDUCDEC	.0162582	.0040332	4.03	0.000	.0083533	.024163
AGE	.0025943	.0011128	2.33	0.020	.0004133	.0047753
FEMALE	.3672884	.0285724	12.85	0.000	.3112876	.4232892
CHILD	.3060317	.0428976	7.13	0.000	.221954	.3901095
FEMCHILD	-.3755503	.0447039	-8.40	0.000	-.4631682	-.2879323
BLACK	-.7104372	.0359462	-19.76	0.000	-.7808903	-.639984
_cons	-.2069298	.1130753	-1.83	0.067	-.4285533	.0146938
/lnalpha	.1674206	.0187562			.1306591	.2041821
alpha	1.182251	.0221746			1.139579	1.226522

margins, dydx(*)

Average marginal effects

Number of obs

=

20,186

Model VCE : Robust

Expression : Predicted number of events, predict()

dy/dx w.r.t. : LC IDP LPI FMDE PHYSLIM NDISEASE HLTHG HLTHF HLTHP LINC LFAM EDU
FEMCHILD BLACK

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
LC	-.1453519	.0450831	-3.22	0.001	-.2337131	-.0569907
IDP	-.425325	.0875932	-4.86	0.000	-.5970046	-.2536454
LPI	.0456313	.012788	3.57	0.000	.0205672	.0706954
FMDE	-.0614799	.0261426	-2.35	0.019	-.1127184	-.0102415
PHYSLIM	.7929485	.0993981	7.98	0.000	.5981318	.9877653
NDISEASE	.0747362	.0050277	14.86	0.000	.0648821	.0845904
HLTHG	.0188377	.0686062	0.27	0.784	-.1156279	.1533033
HLTHF	.6825604	.1261291	5.41	0.000	.435352	.9297689
HLTHP	1.226593	.1973753	6.21	0.000	.8397449	1.613442
LINC	.2435473	.0329971	7.38	0.000	.1788742	.3082204
LFAM	-.3535108	.066902	-5.28	0.000	-.4846363	-.2223852
EDUCDEC	.0468504	.0116396	4.03	0.000	.0240372	.0696636
AGE	.0074759	.0032125	2.33	0.020	.0011796	.0137723
FEMALE	1.058397	.0833728	12.69	0.000	.8949896	1.221805
CHILD	.8818771	.1245455	7.08	0.000	.6377724	1.125982
FEMCHILD	-1.082205	.1297921	-8.34	0.000	-1.336593	-.8278175
BLACK	-2.047233	.1066047	-19.20	0.000	-2.256174	-1.838292

countfit MDU \$XLIST

PRM: Predicted and actual probabilities				
Count	Actual	Predicted	Diff	Pearson
0	0.312	0.107	0.206	8002.381
1	0.189	0.192	0.003	1.111
2	0.138	0.209	0.071	483.374
3	0.093	0.176	0.083	788.603
4	0.067	0.126	0.059	562.277
5	0.048	0.080	0.032	258.557
6	0.034	0.047	0.013	70.960
7	0.026	0.026	0.000	0.009
8	0.020	0.015	0.006	42.205
9	0.014	0.008	0.006	86.069
Sum	0.943	0.987	0.478	1.0e+04
NBRM: Predicted and actual probabilities				
Count	Actual	Predicted	Diff	Pearson
0	0.312	0.318	0.006	2.115
1	0.189	0.191	0.002	0.316
2	0.138	0.128	0.011	18.819
3	0.093	0.089	0.004	3.684
4	0.067	0.064	0.002	1.695
5	0.048	0.047	0.001	0.200
6	0.034	0.035	0.001	0.786
7	0.026	0.027	0.000	0.128
8	0.020	0.020	0.000	0.053
9	0.014	0.016	0.002	3.187
Sum	0.943	0.936	0.029	30.984