

3er Encuentro de Usuarios de Stata en México

**UNA COMPARACIÓN DE LOS MODELOS POISSON
Y BINOMIAL NEGATIVA CON STATA: UN EJERCICIO
DIDÁCTICO**

Noé Becerra Rodríguez

Fortino Vela Peón

Mayo, 2011

Motivación

Actividad docente en los temas de econometría a nivel licenciatura y posgrado.

descriptiva.wordpress.com

einferencial.wordpress.com

mregresion.wordpress.com

tregresion.wordpress.com

Modelos más realistas a situaciones que se presentan en diferentes campos disciplinarios.

Forma sencilla de temas avanzados.

Modelos de variable dependiente limitada

Admiten trabajar con variables dependientes con un rango restringido de valores (binarias con valores 0 y 1, valores enteros, etc.).

- Elección binaria.
- Elección discreta.
- Elección múltiple.
- Datos de recuento.
- Tobit.
- Censurado.
- Truncado.

Modelo de datos de recuento

Aquel que tiene como **variable dependiente** una variable discreta **de conteo** que toma **valores no negativos**.

Modelos de regresión Poisson.

Modelos de regresión binomial negativa.

Modelos de regresión exponencial.

Los modelos de datos de conteo se **caracterizan porque no tienen un límite superior natural**, toman valor cero (en un porcentaje no despreciable) para algunos miembros de la población y suelen tomar pocos valores.

El objetivo consiste en **modelar la distribución de Y_i dado un conjunto de características** eligiendo formas funcionales que aseguren valores positivos.

Modelo de regresión Poisson

La variable **Y** toma pocos valores.

Modelar la distribución de Y_i dado X asumiendo que **Y** dado X_1, X_2, \dots, X_k sigue una **distribución Poisson**, esto es,

$$p[Y_i = Y \setminus X] = \frac{\exp^{-\lambda_i} \lambda_i^y}{y!}$$

o bien, el **valor esperado de Y_i dado X** , esto es

$$E[Y_i = Y \setminus X]$$

La **distribución Poisson** viene determinada **completamente por su media** (todas las probabilidades y momentos de orden superior están determinados por la media).

Esto **impone la restricción** $E(Y|X) = V(Y|X)$, la cual no siempre se cumple en las aplicaciones empíricas.

El **método** de estimación a seguir es el de **máxima verosimilitud (MV)** que podría ofrecer estimadores inconsistentes si la función de probabilidad no está bien especificada.

No obstante, se pueden obtener estimadores consistentes y asintóticamente normales de las β_j si la media condicional esta bien especificada.

Cuando Y dado X_1, X_2, \dots, X_k no sigue una distribución Poisson, el estimador que se obtiene de maximizar el logaritmo de la función de verosimilitud, $L(\beta)$, se le denomina **estimador de cuasi máxima verosimilitud (QML)**.

Cuando se estima por QML si no se cumple el supuesto de $E(Y | \mathbf{X}) = V(Y | \mathbf{X})$ es **necesario ajustar los errores estándar**.

Una posibilidad es **ajustar** considerando que la **varianza es proporcional a la media**, esto es: $V(Y | \mathbf{X}) = \sigma^2 E(Y | \mathbf{X})$, donde σ^2 es un parámetro desconocido.

- Si $\sigma^2 = 1$ **equidispersión**.
- Si $\sigma^2 > 1$ se tiene **sobredispersión** (muy común).
- Si $\sigma^2 < 1$ **infradispersión** (poco común).

Bajo el supuesto de **varianza proporcional a la media** el ajuste de los errores estándar de MV da por resultado a los errores estándar de los modelos lineales generalizados (GML).

Modelo de regresión binomial negativa

El enfoque QML no permite calcular probabilidades condicionales del tipo

$$p(y_i = y \mid x_i) = \frac{e^{-\lambda_i} \lambda_i^y}{y!}$$

Solo se estima

$$E(\mathbf{Y} \mid \mathbf{X})$$

Es necesario considerar modelos alternativos.

Una posibilidad es (Cameron y Trivedi, 1986):

$$V(y_i \mid X_i) = (1 + \delta^2) \exp(X_i \beta)$$

para algún $\delta^2 > 0$ a ser estimado.

Otra es (Cameron y Trivedi, 1986):

$$V(y_i \mid X_i) = (1 + \alpha^2 \exp(X_i \beta)) \exp(X_i \beta)$$

para algún $\alpha^2 > 0$.

Base de datos

```
. set more on, permanent
(set more preference recorded)

. use "C:\Users\Owner\Desktop\base publicaciones.dta", clear

. more

. describe

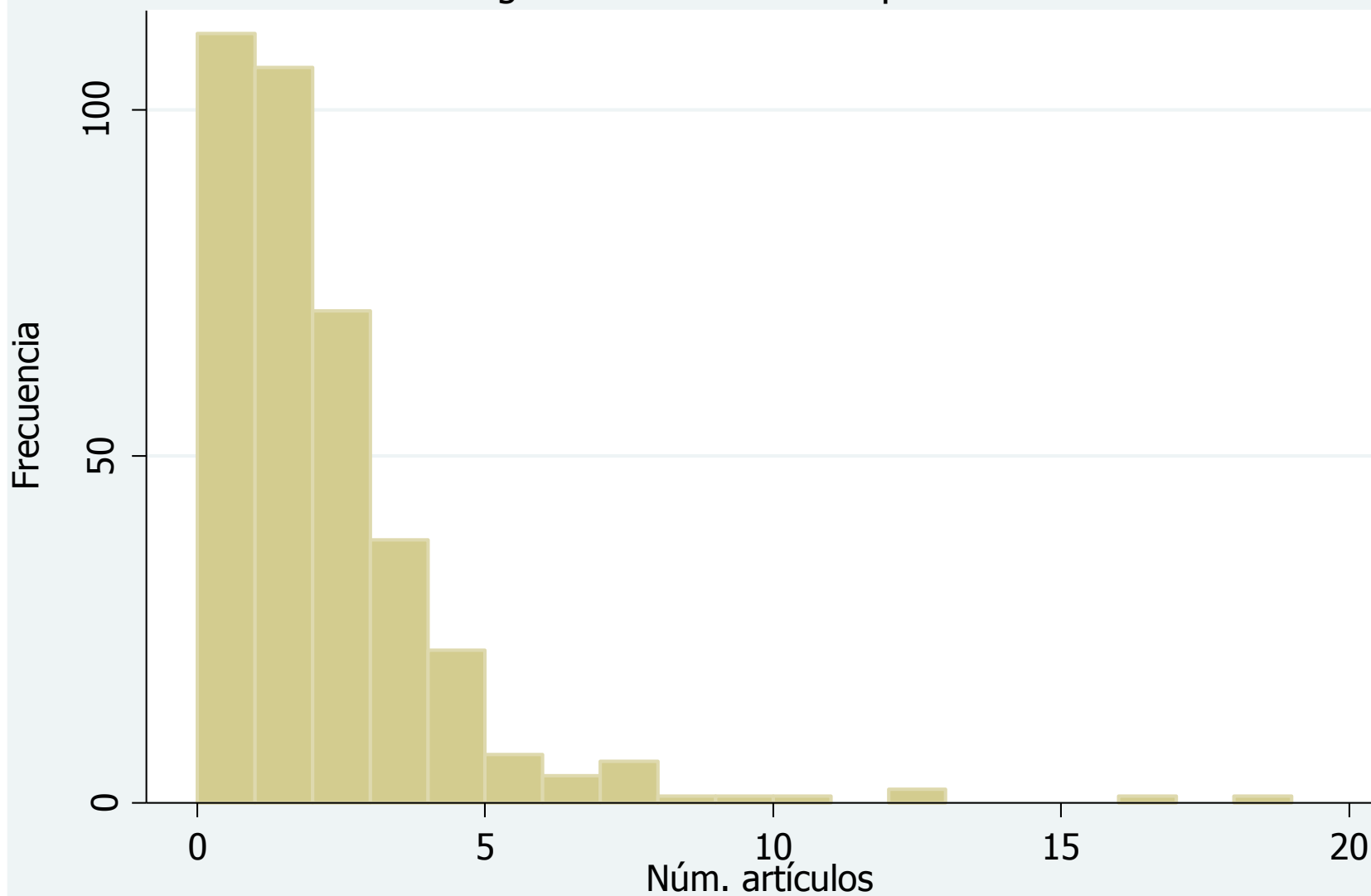
Contains data from C:\Users\Owner\Desktop\base publicaciones.dta
  obs:          372
  vars:          6                               10 May 2011 13:52
  size:         4,092 (99.9% of memory free)
```

variable name	storage type	display format	value label	variable label
publi	byte	%8.0g		Núm. artículos
género	byte	%8.0g		Género investigador
edad	byte	%8.0g		Edad investigador
edad_sqr	int	%8.0g		Cuadrado edad
vinc_empresa	byte	%8.0g		Vínculos con empresas
país_phd	byte	%8.0g		país del PhD

```
Sorted by:

. more
```

Histograma del número de publicaciones



Estadística descriptiva de las publicaciones

```
. summarize publi, detail
```

```
      Núm. artículos
```

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	obs	372
25%	0	0	Sum of wgt.	372
50%	1		Mean	1.712366
		Largest	Std. Dev.	2.175719
75%	2	12		
90%	4	12	Variance	4.733755
95%	5	16	Skewness	3.319603
99%	12	19	Kurtosis	20.62949

```
. more
```

Estimación Poisson

```
. poisson publi género edad edad_sqr vinc_empresa país_phd, nolog
Poisson regression                               Number of obs   =           372
                                                  LR chi2(5)       =           39.39
                                                  Prob > chi2      =           0.0000
Log likelihood = -721.37549                       Pseudo R2       =           0.0266
```

publi	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
género	-.3491275	.0832155	-4.20	0.000	-.5122268 -.1860281
edad	.1703063	.046286	3.68	0.000	.0795873 .2610252
edad_sqr	-.001712	.0004813	-3.56	0.000	-.0026554 -.0007687
vinc_empresa	.1010469	.0833831	1.21	0.226	-.062381 .2644749
país_phd	.1215446	.0816882	1.49	0.137	-.0385613 .2816505
_cons	-3.545044	1.092441	-3.25	0.001	-5.68619 -1.403898

```
. more
. estat gof
Goodness-of-fit chi2 = 767.5677
Prob > chi2(366)    = 0.0000
```


Estimación MLG, familia Poisson y función de enlace Log

```

. more
.
. glm publi género edad edad_sqr vinc_empresa país_phd, family(Poisson) link(log) nolog

Generalized linear models                               No. of obs       =        372
Optimization      : ML                               Residual df      =        366
                                                         Scale parameter  =         1
Deviance          = 767.5677434                       (1/df) Deviance =    2.09718
Pearson          = 888.0953854                       (1/df) Pearson  =    2.42649

Variance function: v(u) = u                           [Poisson]
Link function     : g(u) = ln(u)                       [Log]

Log likelihood    = -721.3754881                       AIC              =    3.910621
                                                         BIC              = -1398.747

```

publi	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
género	-.3491274	.0832155	-4.20	0.000	-.5122268	-.1860281
edad	.1703066	.046286	3.68	0.000	.0795876	.2610256
edad_sqr	-.0017121	.0004813	-3.56	0.000	-.0026554	-.0007687
vinc_empresa	.1010469	.0833831	1.21	0.226	-.0623811	.2644748
país_phd	.1215446	.0816882	1.49	0.137	-.0385613	.2816504
_cons	-3.545051	1.092442	-3.25	0.001	-5.686198	-1.403904

```

. more
.

```

Estimación MLG, fam. Poisson, link log con opción scale(x2)

```
. glm publi género edad edad_sqr vinc_empresa país_phd, family(Poisson) link(log) scale(x2) nolog

Generalized linear models                No. of obs      =       372
Optimization      : ML                  Residual df    =       366
                                                Scale parameter =         1
Deviance          = 767.5677434          (1/df) Deviance = 2.09718
Pearson          = 888.0953854          (1/df) Pearson  = 2.42649

Variance function: V(u) = u                [Poisson]
Link function    : g(u) = ln(u)           [Log]

Log likelihood   = -721.3754881          AIC             = 3.910621
                                                BIC             = -1398.747
```

publi	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
género	-.3491274	.1296264	-2.69	0.007	-.6031904	-.0950645
edad	.1703066	.0721007	2.36	0.018	.0289918	.3116213
edad_sqr	-.0017121	.0007497	-2.28	0.022	-.0031815	-.0002426
vinc_empresa	.1010469	.1298875	0.78	0.437	-.153528	.3556218
país_phd	.1215446	.1272473	0.96	0.339	-.1278555	.3709446
_cons	-3.545051	1.701718	-2.08	0.037	-6.880357	-.2097451

(Standard errors scaled using square root of Pearson X2-based dispersion.)

```
. more
```

Estimación Binomial Negativa

```
. more
. nbreg publi género edad edad_sqr vinc_empresa país_phd, nolog

Negative binomial regression              Number of obs   =           372
                                          LR chi2(5)       =           18.34
Dispersion = mean                        Prob > chi2      =           0.0025
Log likelihood = -651.48568              Pseudo R2       =           0.0139
```

	publi	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
	género	-.3366895	.1176026	-2.86	0.004	-.5671863	-.1061926
	edad	.1683063	.0627778	2.68	0.007	.0452642	.2913485
	edad_sqr	-.0016875	.0006493	-2.60	0.009	-.0029601	-.0004149
	vinc_empresa	.0894518	.1189718	0.75	0.452	-.1437286	.3226323
	país_phd	.101695	.1174399	0.87	0.387	-.1284831	.331873
	_cons	-3.494401	1.487225	-2.35	0.019	-6.409308	-.5794938
	/lnalpha	-.4803214	.1537219			-.7816108	-.1790321
	alpha	.6185845	.09509			.4576682	.8360791

```
Likelihood-ratio test of alpha=0:  chibar2(01) = 139.78 Prob>=chibar2 = 0.000
```

Estimación MLG, familia Binomial Negativa, link log

```

. more

. glm publi género edad edad_sqr vinc_empresa país_phd, family(nbinomial) link(log) nolog

Generalized linear models                               No. of obs       =        372
Optimization      : ML                               Residual df      =        366
                                                         Scale parameter  =         1
Deviance          = 321.8246429                       (1/df) Deviance  =  .8793023
Pearson          = 315.7286879                       (1/df) Pearson   =  .8626467

Variance function: V(u) = u+(1)u^2                   [Neg. Binomial]
Link function     : g(u) = ln(u)                     [Log]

Log likelihood    = -657.0137971                     AIC              =    3.56459
                                                         BIC              = -1844.491

```

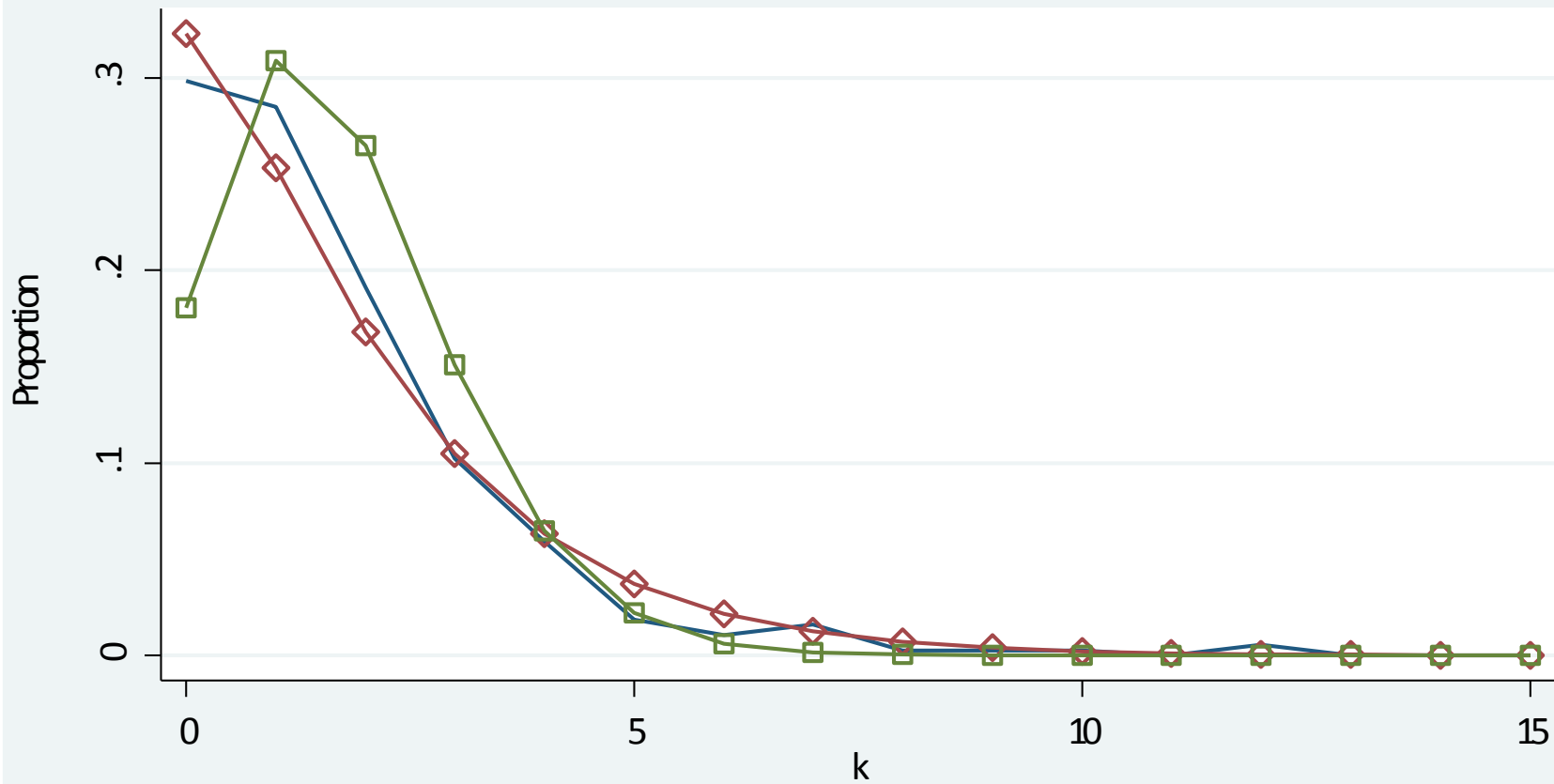
publi	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
género	-.3337564	.1344289	-2.48	0.013	-.5972322	-.0702805
edad	.1676824	.0709026	2.36	0.018	.0287158	.306649
edad_sqr	-.0016793	.0007319	-2.29	0.022	-.0031138	-.0002448
vinc_empresa	.0872908	.1361298	0.64	0.521	-.1795187	.3541004
país_phd	.0974453	.1346417	0.72	0.469	-.1664475	.3613382
_cons	-3.481436	1.68227	-2.07	0.039	-6.778625	-.1842477

```

. more

```

Ajuste Poisson y Binomial Negativa a publicaciones



mean = 1.712; overdispersion = .6901



Estimadores modelos Poisson y Binomial Negativa

	Poisson	Binomial N~a
Núm. artículos		
Género investigador	-0.349*** (0.000)	-0.337** (0.004)
Edad investigador	0.170*** (0.000)	0.168** (0.007)
Cuadrado edad	-0.00171*** (0.000)	-0.00169** (0.009)
vínculos con empre~s	0.101 (0.226)	0.0895 (0.452)
país del PhD	0.122 (0.137)	0.102 (0.387)
Constant	-3.545** (0.001)	-3.494* (0.019)
Inalpha		
Constant		-0.480** (0.002)
observations	372	372

p-values in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Estimadores Modelos Lineales Generalizados

	Poisson	Pois. escala	Bin Negativa	Gamma
Núm. artículos				
Género investigador	-0.349*** (0.0832)	-0.349** (0.130)	-0.334* (0.134)	-0.324** (0.124)
Edad investigador	0.170*** (0.0463)	0.170* (0.0721)	0.168* (0.0709)	0.165** (0.0620)
Cuadrado edad	-0.00171*** (0.000481)	-0.00171* (0.000750)	-0.00168* (0.000732)	-0.00164** (0.000634)
Vínculos con empre~s	0.101 (0.0834)	0.101 (0.130)	0.0873 (0.136)	0.0806 (0.125)
país del PhD	0.122 (0.0817)	0.122 (0.127)	0.0974 (0.135)	0.0834 (0.124)
Constant	-3.545** (1.092)	-3.545* (1.702)	-3.481* (1.682)	-3.429* (1.483)
Observations	372	372	372	372

Standard errors in parentheses

* p<0.05, ** p<0.01, *** p<0.001

. more

more