11) Poisson and Negative Binomial Regression

Vitor Kamada

August 2019

Reference

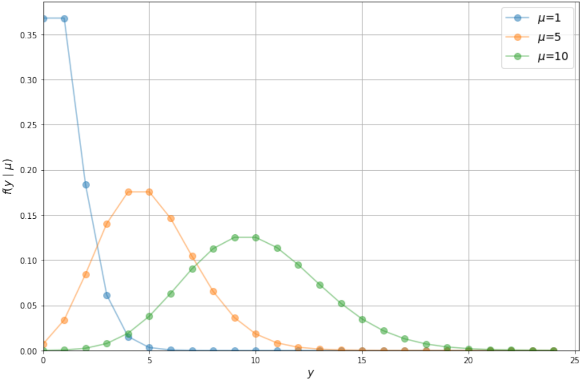
Wooldridge (2010). **Econometric Analysis of Cross Section and Panel Data.** Ch 18.1 to 18.3

https://ebookcentral.proquest.com/lib/wayne/detail.action?docID=3339196&

Poisson Distribution

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

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



$Y \sim$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	

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

Poisson MLE

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

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

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

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

Poisson ML vs Pseudo-ML (PML) or Quasi-ML (QML)

$$V_{PML}[\hat{\beta}_p]$$

$$= (\sum_{i=1}^{N} \mu_i x_i x_i')^{-1} (\sum_{i=1}^{N} w_i x_i x_i') (\sum_{i=1}^{N} \mu_i x_i x_i')^{-1}$$

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

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

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(coinsurance + 1), $0 \leq$ coinsurance ≤ 100	1.710	1.962
IDP	1 if individual deductible plan, 0 otherwise	0.220	0.414
LPI	ln(max(1,annual participation incentive payment))	4.709	2.697
FMDE	0 if IDP = 1	3.153	3.641
	ln(max(1,MDE/(0.01 coinsurance))) otherwise		
LINC	ln(family income)	8.708	1.228
LFAM	ln(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)

Load the Rand Data Set

```
import numpy as np
import statsmodels.api as sm
import pandas as pd
```

```
file="https://github.com/VitorKamada/ECO7110/raw/master/Data/Rand.dta"
df = pd.read_stata(file)
```

```
df['const'] = 1
Xs = ['LC', 'IDP', 'LPI', 'FMDE', 'PHYSLIM','NDISEASE',
      'HLTHG', 'HLTHF','HLTHP','LINC','LFAM','EDUCDEC',
     'AGE','FEMALE','CHILD','FEMCHILD','BLACK','const']
```

```python
Poisson = sm.Poisson(df.MDU, df[Xs]).fit()
print(Poisson.summary())
```

	coef	std err	z	P>|z|
LC	-0.0427	0.006	-7.030	0.000
IDP	-0.1613	0.012	-13.881	0.000
LPI	0.0129	0.002	6.999	0.000
FMDE	-0.0206	0.004	-5.803	0.000
PHYSLIM	0.2684	0.012	21.711	0.000
NDISEASE	0.0232	0.001	38.124	0.000
HLTHG	0.0394	0.010	4.109	0.000
HLTHF	0.2531	0.016	15.613	0.000
HLTHP	0.5216	0.027	19.150	0.000
LINC	0.0834	0.005	16.147	0.000
LFAM	-0.1297	0.009	-14.471	0.000
EDUCDEC	0.0176	0.002	10.749	0.000
AGE	0.0024	0.000	5.510	0.000
FEMALE	0.3488	0.011	30.727	0.000
CHILD	0.3362	0.018	18.866	0.000
FEMCHILD	-0.3625	0.018	-20.208	0.000
BLACK	-0.6801	0.016	-43.738	0.000

```python
PoissonRobust = sm.Poisson(df.MDU,
      df[Xs]).fit(cov_type='HC1')
```

	coef	std err	z	P>|z|
LC	-0.0427	0.015	-2.835	0.005
IDP	-0.1613	0.028	-5.773	0.000
LPI	0.0129	0.004	2.912	0.004
FMDE	-0.0206	0.009	-2.319	0.020
PHYSLIM	0.2684	0.033	8.240	0.000
NDISEASE	0.0232	0.002	13.487	0.000
HLTHG	0.0394	0.023	1.699	0.089
HLTHF	0.2531	0.043	5.894	0.000
HLTHP	0.5216	0.075	6.966	0.000
LINC	0.0834	0.014	5.993	0.000
LFAM	-0.1297	0.023	-5.717	0.000
EDUCDEC	0.0176	0.004	4.358	0.000
AGE	0.0024	0.001	2.124	0.034
FEMALE	0.3488	0.028	12.300	0.000
CHILD	0.3362	0.040	8.319	0.000
FEMCHILD	-0.3625	0.044	-8.211	0.000
BLACK	-0.6801	0.037	-18.443	0.000

```
PoissonRobust_margeff = PoissonRobust.get_margeff()
print(PoissonRobust.summary())
```

```
                  coef    std err          z      P>|z|
-----------------------------------------------------------
LC             -0.0427      0.015     -2.835      0.005
IDP            -0.1613      0.028     -5.773      0.000
LPI             0.0129      0.004      2.912      0.004
FMDE           -0.0206      0.009     -2.319      0.020
PHYSLIM         0.2684      0.033      8.240      0.000
NDISEASE        0.0232      0.002     13.487      0.000
HLTHG           0.0394      0.023      1.699      0.089
HLTHF           0.2531      0.043      5.894      0.000
HLTHP           0.5216      0.075      6.966      0.000
LINC            0.0834      0.014      5.993      0.000
LFAM           -0.1297      0.023     -5.717      0.000
EDUCDEC         0.0176      0.004      4.358      0.000
AGE             0.0024      0.001      2.124      0.034
FEMALE          0.3488      0.028     12.300      0.000
CHILD           0.3362      0.040      8.319      0.000
FEMCHILD       -0.3625      0.044     -8.211      0.000
BLACK          -0.6801      0.037    -18.443      0.000
```

Negative Binomial Distribution

$$Pr(Y = y|\mu, \alpha)$$

$$= \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}$$

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

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

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

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

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

$$\boxed{\sum_{i=1}^{N} (\tfrac{y_i-\mu_i}{1+\alpha\mu_i})x_i = 0}$$

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

Overdispersion

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

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

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

$$H_0 : \alpha = 0 \text{ } \textbf{(equidispersion)}$$

```
NB1 = sm.NegativeBinomial(df.MDU, df[Xs],
          loglike_method='nb1').fit()
```

	coef	std err	z	P>|z|
LC	-0.0569	0.011	-5.208	0.000
IDP	-0.1802	0.021	-8.682	0.000
LPI	0.0143	0.003	4.323	0.000
FMDE	-0.0135	0.006	-2.121	0.034
PHYSLIM	0.1977	0.024	8.407	0.000
NDISEASE	0.0200	0.001	17.520	0.000
HLTHG	0.0390	0.017	2.283	0.022
HLTHF	0.2145	0.031	7.009	0.000
HLTHP	0.5208	0.054	9.617	0.000
LINC	0.0739	0.009	7.970	0.000
LFAM	-0.0959	0.016	-5.921	0.000
EDUCDEC	0.0220	0.003	7.389	0.000
AGE	0.0017	0.001	2.162	0.031
FEMALE	0.3705	0.021	18.039	0.000
CHILD	0.3204	0.031	10.178	0.000
FEMCHILD	-0.3852	0.032	-12.055	0.000
BLACK	-0.7237	0.028	-26.206	0.000
const	-0.1190	0.089	-1.344	0.179
alpha	3.4608	0.059	58.858	0.000

```
NB1_margeff = NB1.get_margeff()
```

	dy/dx	std err	z	P>|z|
LC	-0.1628	0.031	-5.203	0.000
IDP	-0.5156	0.060	-8.657	0.000
LPI	0.0409	0.009	4.319	0.000
FMDE	-0.0388	0.018	-2.121	0.034
PHYSLIM	0.5657	0.067	8.385	0.000
NDISEASE	0.0572	0.003	17.316	0.000
HLTHG	0.1116	0.049	2.282	0.022
HLTHF	0.6138	0.088	6.996	0.000
HLTHP	1.4902	0.156	9.582	0.000
LINC	0.2114	0.027	7.950	0.000
LFAM	-0.2744	0.046	-5.913	0.000
EDUCDEC	0.0630	0.009	7.373	0.000
AGE	0.0049	0.002	2.161	0.031
FEMALE	1.0599	0.059	17.816	0.000
CHILD	0.9166	0.090	10.138	0.000
FEMCHILD	-1.1021	0.092	-11.988	0.000
BLACK	-2.0707	0.081	-25.537	0.000

```python
NB2 = sm.NegativeBinomial(df.MDU, df[Xs]).fit()
print(NB2.summary())
```

	coef	std err	z	P>|z|
LC	-0.0504	0.013	-3.913	0.000
IDP	-0.1476	0.025	-5.810	0.000
LPI	0.0158	0.004	3.897	0.000
FMDE	-0.0214	0.008	-2.844	0.004
PHYSLIM	0.2746	0.030	9.292	0.000
NDISEASE	0.0259	0.001	17.501	0.000
HLTHG	0.0067	0.020	0.331	0.740
HLTHF	0.2370	0.037	6.337	0.000
HLTHP	0.4266	0.074	5.750	0.000
LINC	0.0847	0.009	9.890	0.000
LFAM	-0.1226	0.019	-6.348	0.000
EDUCDEC	0.0163	0.003	4.680	0.000
AGE	0.0026	0.001	2.759	0.006
FEMALE	0.3667	0.024	15.277	0.000
CHILD	0.3058	0.039	7.930	0.000
FEMCHILD	-0.3746	0.037	-10.087	0.000
BLACK	-0.7104	0.027	-25.841	0.000
const	-0.2095	0.090	-2.330	0.020
alpha	1.1822	0.017	67.613	0.000

```
NB2_margeff = NB2.get_margeff()
```

	dy/dx	std err	z	P>|z|
LC	-0.1451	0.037	-3.906	0.000
IDP	-0.4254	0.073	-5.791	0.000
LPI	0.0456	0.012	3.889	0.000
FMDE	-0.0616	0.022	-2.841	0.004
PHYSLIM	0.7913	0.086	9.183	0.000
NDISEASE	0.0748	0.004	16.797	0.000
HLTHG	0.0193	0.058	0.331	0.740
HLTHF	0.6830	0.108	6.309	0.000
HLTHP	1.2292	0.215	5.728	0.000
LINC	0.2441	0.025	9.803	0.000
LFAM	-0.3531	0.056	-6.321	0.000
EDUCDEC	0.0470	0.010	4.670	0.000
AGE	0.0075	0.003	2.757	0.006
FEMALE	1.0567	0.071	14.892	0.000
CHILD	0.8811	0.112	7.884	0.000
FEMCHILD	-1.0795	0.108	-9.975	0.000
BLACK	-2.0471	0.084	-24.336	0.000