Skellam regression in Stata

Vincenzo Verardi
FNRS, U. Namur (DeFIPP-CRED), ULB
joint with Catherine Vermandele

UK Stata Conference
(London-LSE)

September 8, 2023

Introduction

The **Skellam distribution** is a probability distribution that models the **difference between two Poisson** random variables that are independent of one another and that might have **different means**.

It is named after **British statistician** and ecologist **John Gordon Skellam** (1914-1979).

It is a **generalisation of Irwin distribution** (see Irwin, 1937) that models the **difference between two independent Poisson** random variables that share the **same mean**.

A **Skellam regression** uses **Maximum Likelihood** to estimate how the **conditional means** of the underlying poisson processes are **related to a set of covariates**.

In this talk, we show how to **write the ML problem** and get the **gradient** and **Hessian** for numerical optimization. A simple Stata implementation is presented.

Remark: These slides are the **support** for an **oral presentation** were **additionnal information** is **provided**. They should **not** be considered as **a stand alone document**.

Some examples in the literature

Kendall (1951) and Dobbie (1961) show that the Skellam distribution can be used in the **problem of taxis and customers coming to a waiting area** in different Poisson flows (i.e. with different rates). The number of **taxis waiting** is the (integer) **variable of interest**. This number can be **positive if taxis are waiting**, **zero if there is no queue**, or **negative** if **customers are waiting**.

It is also often used to model the **number of points that separate two teams in sports** such as hockey and soccer. See, for example, Karlis and Ntzoufras (2008).

More recently, Liu and Pelechrinis (2021) look at the case of **shared transportation**. They use a Skellam regression to predict the **difference in overall demand and supply** at a particular bike station over a certain time period.

Modified Bessel function of the first kind

The **Modified Bessel Function of the First Kind** arises in many areas of mathematics and physics. It is denoted by $I_k(x)$. We are only interested in the case where order $k \in \mathbb{Z}$ and $x \in \mathbb{R}^+$here. It is defined as:

$$I_k(x) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+k+1)} \left(\frac{x}{2}\right)^{2m+k},$$

where $\Gamma(\cdot)$ is the Gamma function $\left(\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt\right)$.

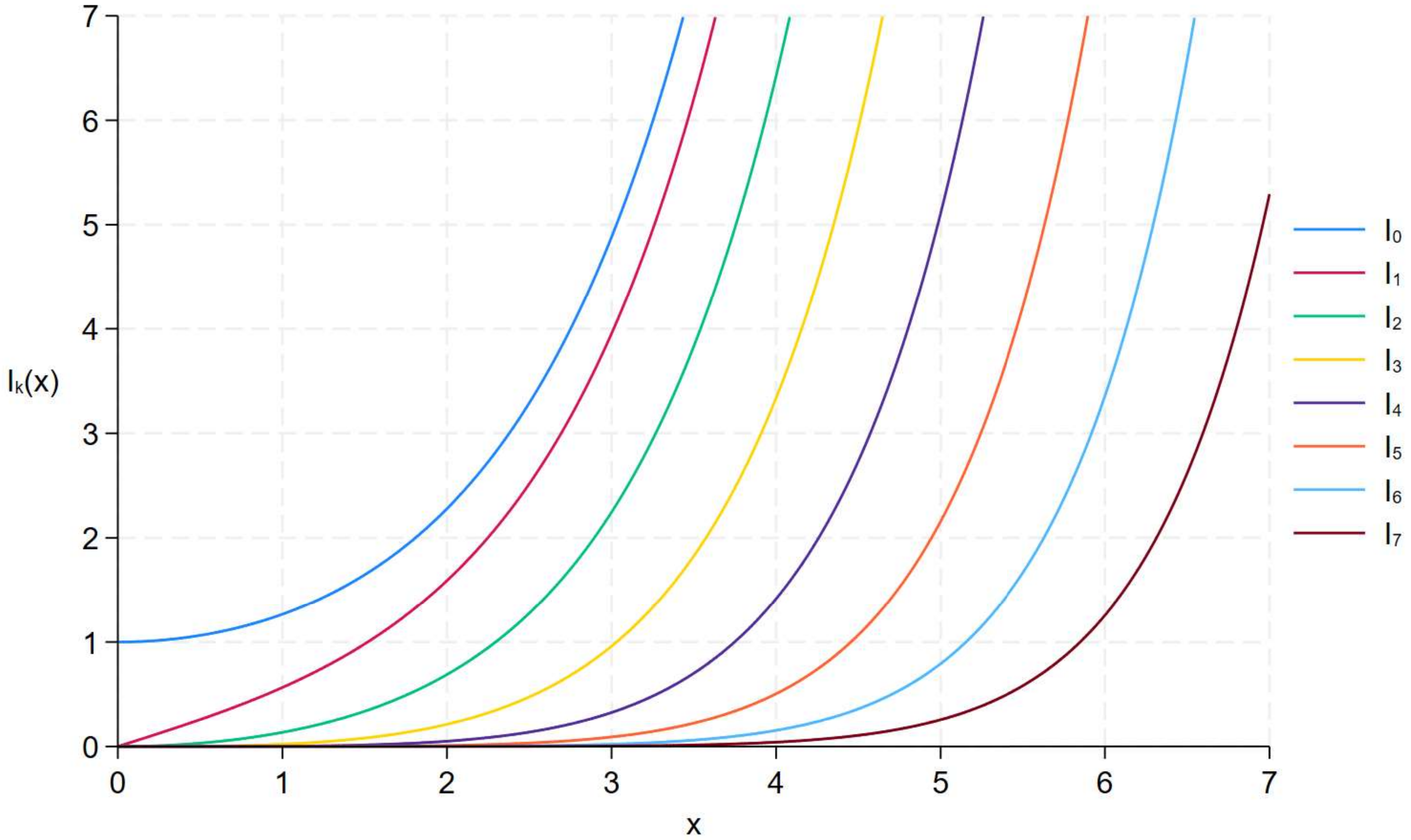
If the **values of** k **are integers** (as in our case), $I_{-k}(x) = I_k(x)$ (see Abramowitz and Stegun 1972, p. 375, 9.6.6). $I_k(\cdot)$ can thus be replaced by $I_{|k|}(\cdot)$ in the above formula.

Furthermore (see Abramowitz and Stegun 1972, p.376, 9.6.26), for $k \in \mathbb{Z}$,

$$I_k'(z) = \frac{d}{dz} I_k(z) = \frac{I_{k-1}(z) + I_{k+1}(z)}{2}$$

To the best of our knowledge this **function is not available in Stata**. However, we have adapted the C++ code by Moreau (2011), with permission, to be compatible with Mata.

Modified Bessel function of the first kind



Skellam distribution

Let Y_1 and Y_2 be **two independent Poisson-distributed random variables** with means μ_1 and μ_2. Then, $Y = Y_1 - Y_2$ has a Skellam distribution. Its **probability mass function** is given by

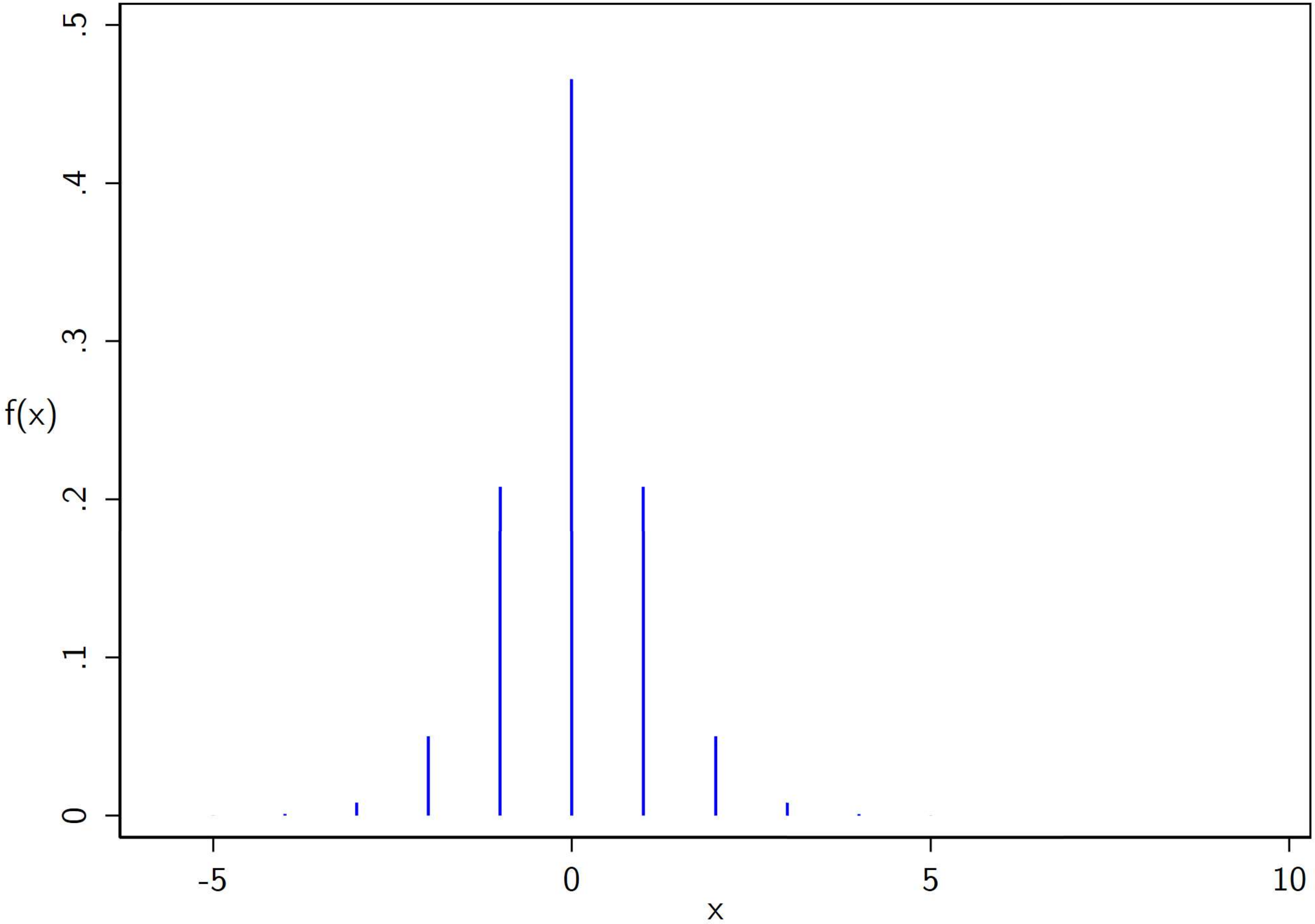
$$\Pr\{Y = k\} = e^{-(\mu_1+\mu_2)} \left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_{|k|}(2\sqrt{\mu_1\mu_2})$$

where $k \in \mathbb{Z}$ and where $I_k(\cdot)$ is the modified Bessel function of the first kind.

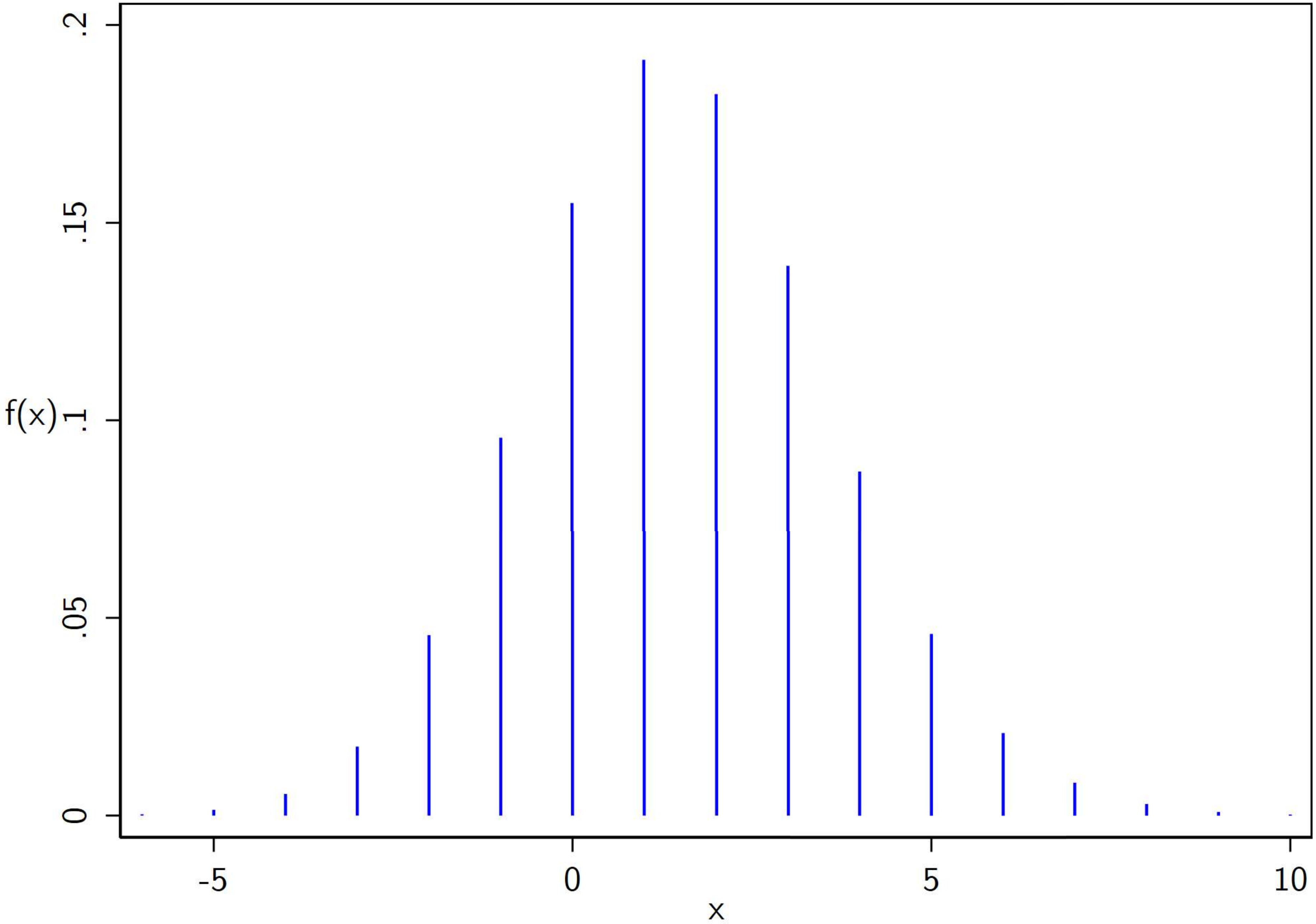
To **guarantee positiveness** of μ_1 and μ_2, the probability mass function can be **reparametrized** by defining $\mu_1 = \exp(\lambda_1)$ and $\mu_2 = \exp(\lambda_2)$ and can be re-written as

$$\Pr\{Y = k\} = e^{-\left(e^{\lambda_1}+e^{\lambda_2}\right)} \left(e^{\lambda_1-\lambda_2}\right)^{k/2} I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)$$

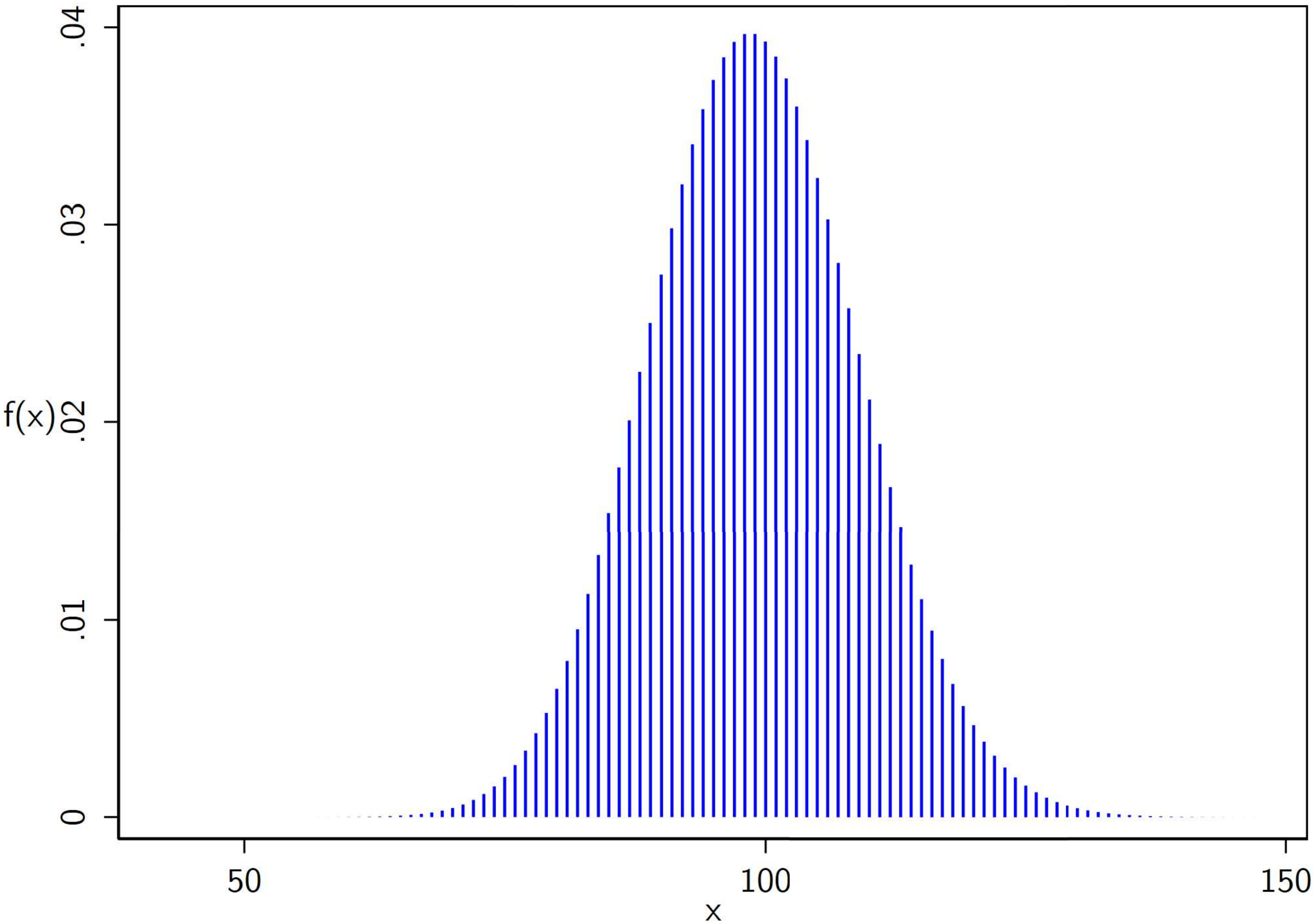
Skellam(μ_1 =0.5,μ_2 =0.5)



Skellam(μ_1 =3,μ_2 =1.5)



Skellam($\mu_1 = 100, \mu_2 = 1$)



Maximum Likelihood estimator

The likelihood function is given by

$$\mathscr{L}(\lambda_1, \lambda_2; k_1, \dots, k_n) = \prod_{i=1}^{n} \Pr(Y_i = k_i \mid \lambda_1, \lambda_2)$$

$$= \prod_{i=1}^{n} \left\{ e^{-\left(e^{\lambda_1} + e^{\lambda_2}\right)} \left(e^{\lambda_1 - \lambda_2}\right)^{\frac{k_i}{2}} I_{|k_i|}\left(2\sqrt{e^{\lambda_1 + \lambda_2}}\right) \right\}$$

The maximum likelihood estimates $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ of the two parameters of the Skellam distribution are solutions of the maximization problem

$$\max_{\lambda_1, \lambda_2 \in \mathbb{R}} \ln \mathscr{L}(\lambda_1, \lambda_2; k_1, \dots, k_n) = \max_{\lambda_1, \lambda_2 \in \mathbb{R}} \sum_{i=1}^{n} L(\lambda_1, \lambda_2; k_i)$$

where

$$L(\lambda_1, \lambda_2; k) = -\left(e^{\lambda_1} + e^{\lambda_2}\right) + (\lambda_1 - \lambda_2)\frac{k_i}{2} + \ln I_{|k|}\left(2\sqrt{e^{\lambda_1 + \lambda_2}}\right), \quad k \in \mathbb{Z}$$

Some available implementations

The maximum likelihood estimates $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ of the two parameters of the Skellam distribution are solutions of the maximization problem

$$\max_{\lambda_1,\lambda_2\in\mathbb{R}} \ln \mathscr{L}(\lambda_1,\lambda_2;k_1,\ldots,k_n) = \max_{\lambda_1,\lambda_2\in\mathbb{R}} \sum_{i=1}^{n} L(\lambda_1,\lambda_2;k_i)$$

where

$$L(\lambda_1,\lambda_2;k) = -\left(e^{\lambda_1}+e^{\lambda_2}\right) + (\lambda_1-\lambda_2)\frac{k_i}{2} + \ln I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right), \quad k\in\mathbb{Z}$$

R: by Michail Tsagris (See Lewis, Brown and Tsagris, 2017).

- Newton-type `nlm`
- Nelder-Mead is used by `optim` after `nlm` to improve the preliminary result

Python: by Liu and Pelechrinis (2021)

- Optimization is done relying on the Conjugate Gradient method

Maximum Likelihood estimator

To solve this maximisation problem, the gradient and the Hessian, with respect to λ_1 and λ_2, of the log-likelihood function, and hence of function $L(\lambda_1, \lambda_2; k)$, can be computed. Since, for $k \in \mathbb{Z}$,

$$I'_k(z) = \frac{d}{dz} I_k(z) = \frac{I_{k-1}(z) + I_{k+1}(z)}{2}$$

(see 9.6.26 page 376 in Abramowitz and Stegun, 1972), we have the following first derivatives for the gradient:

$$\frac{\partial}{\partial \lambda_1} L(\lambda_1, \lambda_2; k) = -e^{\lambda_1} + \frac{k}{2} + \frac{\sqrt{e^{\lambda_1+\lambda_2}}}{2} \left[\frac{I_{||k|-1|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right) + I_{|k|+1}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}{I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)} \right]$$

$$\frac{\partial}{\partial \lambda_2} L(\lambda_1, \lambda_2; k) = -e^{\lambda_2} - \frac{k}{2} + \frac{\sqrt{e^{\lambda_1+\lambda_2}}}{2} \left[\frac{I_{||k|-1|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right) + I_{|k|+1}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}{I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)} \right]$$

Maximum Likelihood estimator

For the Hessian, let's first calculate the cross derivatives:

$$
\begin{aligned}
\frac{\partial^2}{\partial\lambda_1\partial\lambda_2} L(\lambda_1,\lambda_2;k) &= \frac{\partial^2}{\partial\lambda_2\partial\lambda_1} L(\lambda_1,\lambda_2;k) \\
&= \frac{e^{\lambda_1+\lambda_2}}{2} + \frac{e^{\lambda_1+\lambda_2}}{4}\left[\frac{I_{||k|-2|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)+I_{|k|+2}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}{I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}\right] \\
&+ \frac{\sqrt{e^{\lambda_1+\lambda_2}}}{4}\left[\frac{I_{||k|-1|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)+I_{|k|+1}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}{I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}\right] \\
&\times \left\{1-\sqrt{e^{\lambda_1+\lambda_2}}\left[\frac{I_{||k|-1|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)+I_{|k|+1}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}{I_{|k|}\left(2\sqrt{e^{\lambda_1+\lambda_2}}\right)}\right]\right\}
\end{aligned}
$$

The second derivatives are given by

$$\frac{\partial^2}{\partial\lambda_1^2} L(\lambda_1,\lambda_2;k) = -e^{\lambda_1} + \frac{\partial^2}{\partial\lambda_1\partial\lambda_2} L(\lambda_1,\lambda_2;k)$$

$$\frac{\partial^2}{\partial\lambda_2^2} L(\lambda_1,\lambda_2;k) = -e^{\lambda_2} + \frac{\partial^2}{\partial\lambda_1\partial\lambda_2} L(\lambda_1,\lambda_2;k)$$

Maximum Likelihood estimator

In the context of **Skellam regression**, the parameters λ_1 and λ_2 of the two independent Poisson distributions are expressed as linear functions of p covariates $X_1, \ldots, X_p$. That is to say that, for $i = 1, \ldots, n$,

$$\Pr\{Y_i = k_i\} = e^{-\left(e^{\lambda_{1i}} + e^{\lambda_{2i}}\right)} \left(e^{\lambda_{1i} - \lambda_{2i}}\right)^{k_i/2} I_{|k_i|}\left(2\sqrt{e^{\lambda_{1i} + \lambda_{2i}}}\right)$$

where $\lambda_{1i} = \mathbf{x}_i^T \beta$ and $\lambda_{2i} = \mathbf{x}_i^T \gamma$, with $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$. We have here to estimate two $(p+1)$-vectors of parameters (β and γ) by solving the maximisation problem

$$\max_{\beta, \gamma \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} L(\beta, \gamma; k_i, \mathbf{x}_i)$$

where, for $i = 1, \ldots, n$

$$L(\beta, \gamma; k_i, \mathbf{x}_i) = -\left(e^{\mathbf{x}_i^T \beta} + e^{\mathbf{x}_i^T \gamma}\right) + \left(\mathbf{x}_i^T \beta - \mathbf{x}_i^T \gamma\right) \frac{k_i}{2} + \ln I_{|k_i|}\left(2\sqrt{e^{\mathbf{x}_i^T \beta + \mathbf{x}_i^T \gamma}}\right)$$

The **first and second derivatives** presented in the previous slide have to be modified and **multiplied** respectively **by** x_i^T **for the gradient** and $\mathbf{x}_i \mathbf{x}_i^T$ **for the second and cross derivatives**.

Stata implementation

Title

skelreg — Skellam regression estimator

Syntax

skelreg *varlist* [*if*] [*in*] **,** [*options*]

To be used with caution as the model is non-linear

options	Description
robust	use the sandwich variance formula to compute standard errors of the estimated parameters.
cluster(*varname*)	compute cluster-corrected standard errors of the estimated parameters.
nolog	do not show iteration logs
noconstant	fit a model without a constant
stub(*string*)	provide a stub for the dependent variable.[1]
technique(*string*)	change optimization technique. See [M-5] optimize##i_technique.[2]
nodofcorrection	do not correct for the degress of freedom
level(*cilevel*)	set the confidence level

[1] The code creates two temporary variables automatically by taking the name of the dependent variable and adding " count 1" and " count 2". If a different name needs to be used (for example, if a variable with the same name already exists in the dataset), the stub option can be used to declare it.

[2] Note that the Nelder-Mead optimization technique is not available here.

Options for predict post-estimation command

ndiff	generates predicted difference in counts between processes (default).
xb1	generates the linear predictions for the first process.
xb2	generates the linear predictions for the second process.
n1	generates predicted counts (i.e. exp(xb1)) for the first process.
n2	generates predicted counts (i.e. exp(xb2)) for the second process.

Description

skelreg The dependent variable in Skellam regression is the difference between two counts, while the explanatory variables are predictors that may affect event frequency.

Simulations

To illustrate how a simple `Stata/Mata` code can be used to estimate the parameters of the Skellam distribution, we first generate $\boldsymbol{n = 1000}$ observations from a **random variable** Y defined as the **difference** $(Y_1 - Y_2)$ **of two independent Poisson**-distributed variables, $Y_1 \sim \mathscr{P}\left(\mu_1 = e^{\lambda_1}\right)$ and $Y_2 \sim \mathscr{P}\left(\mu_2 = e^{\lambda_2}\right)$.

To have an idea of the performance of the estimator, we run some **Monte Carlo simulations** by simply replicating **B=1000 times** this setup. We take $\lambda_1 = 1.3$ **and** $\lambda_2 = 0.7$

j	1	2
λ_j	1.3	0.7
ave $\left\{\widehat{\lambda}_j^{(b)}\right\}$	1.2982	0.6950

j	1	2
s.d. $\left\{\widehat{\lambda}_j^{(b)}\right\}$	0.0376	0.0657
ave $\left\{\text{s.e.}(\widehat{\lambda}_j^{(b)})\right\}$	0.0375	0.0667

Simulations

In a second setup, we change the data generating process and make λ_1 and λ_2 dependent on an explanatory variable X. We use a standard **normal** distribution to **generate** $n = 1000$ **observations** x_i.

We then generate $n = 1000$ observations y_{i1} **from a Poisson** distribution with **mean** $e^{\lambda_{i1}}$ where $\lambda_{i1} = \beta_0 + \beta_1 x_i = 0 + 1.3x_i$, and $n = 1000$ observations y_{i2} from a **Poisson** distribution with mean $\boldsymbol{e}^{\lambda_{i2}}$ where $\lambda_{i2} = \gamma_0 + \gamma_1 x_i = 0 + 0.7x_i$.

Finally, we determine the **observations** $y_i = y_{i1} - y_{i2}$ for $i = 1, \ldots, n$.

As before, we run some **Monte Carlo simulations** by simply replicating B=1000 times this setup.

ℓ	0	1
β_ℓ	0	1.3
ave $\left\{\widehat{\beta}_\ell^{(b)}\right\}$	−0.0040	1.3012
s.d. $\left\{\widehat{\beta}_\ell^{(b)}\right\}$	0.0567	0.0342
ave $\left\{\text{s.e.}(\widehat{\beta}_\ell^{(b)})\right\}$	0.0575	0.0353

ℓ	0	1
γ_ℓ	0	0.7
ave $\left\{\widehat{\gamma}_\ell^{(b)}\right\}$	−0.0050	0.6977
s.d. $\left\{\widehat{\gamma}_\ell^{(b)}\right\}$	0.0573	0.0652
ave $\left\{\text{s.e.}(\widehat{\gamma}_\ell^{(b)})\right\}$	0.0587	0.0652

Synthetic data example

```
clear
program drop _all
set obs 250
gen x=rnormal(2,1)
gen y1=rpoisson(exp(0.6*x))
gen y2=rpoisson(exp(0.4*x))

gen y=y1-y2
skelreg y x*, nolog

test [y_count_1]:x=[y_count_2]:x=0

margins, dydx(*)
predict yhat, ndiff
predict yhatp1, n1
predict yhatp2, n2
```

Synthetic data example

```
. skelreg y x*, nolog
                                                    Number of obs =     250

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
y_count_1    |
           x |   .6389506   .0752313     8.49   0.000        .4915    .7864012
       _cons |   .0055461   .1840548     0.03   0.976    -.3551947    .3662868
-------------+----------------------------------------------------------------
y_count_2    |
           x |   .5019903   .1104102     4.55   0.000     .2855903    .7183902
       _cons |  -.0193426   .2392158    -0.08   0.936    -.4881969    .4495117
------------------------------------------------------------------------------

. test [y_count_1]:x=[y_count_2]:x=0

 ( 1)  [y_count_1]x - [y_count_2]x = 0
 ( 2)  [y_count_1]x = 0

           chi2(  2) =   93.54
         Prob > chi2 =    0.0000
```

Synthetic data example

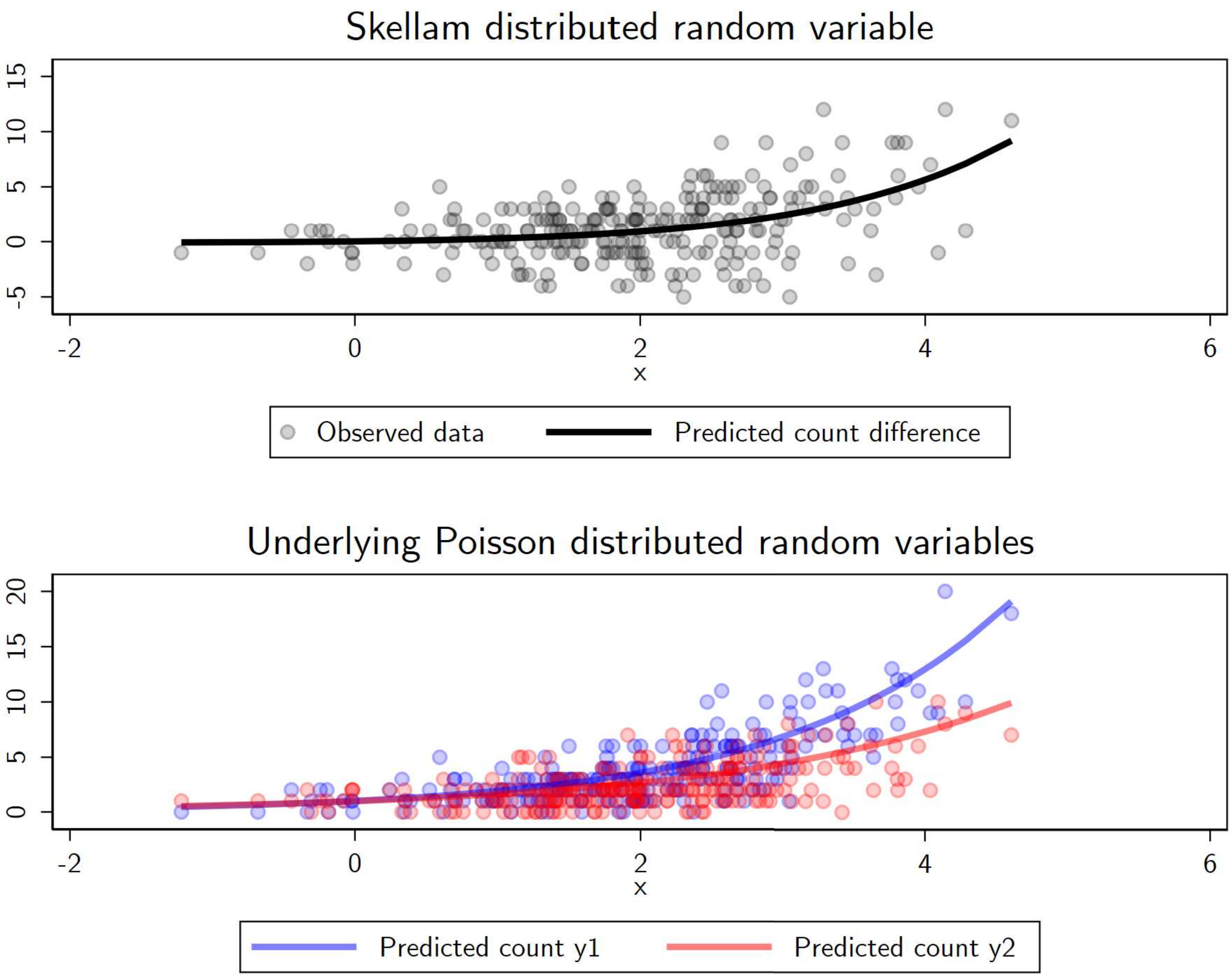
```
. margins, dydx(*)

Average marginal effects                          Number of obs = 250

Expression: predict()
dy/dx wrt:  x
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
x	1.256896	.207978	6.04	0.000	.8492665	1.664525

Synthetic data example



References

Abramowitz, M., and I. A. Stegun, ed. 1972. Handbook of mathematical functions with formulas, graphs, and mathematical tables. National Bureau of Standards Applied Mathematics Series - 55. Issued June 1964. Tenth Printing, December 1972, with corrections.

Dobbie, J. M. 1961. Letter to the Editor — A Doubled-Ended Queuing Problem of Kendall. Operations Research 9(5): 755–757.

Irwin, J. O. 1937. The Frequency Distribution of the Difference between Two Independent Variates following the same Poisson Distribution. Journal of the Royal Statistical Society (Series A) 100(3): 415–416

Karlis D. and I. Ntzoufras. 2009. Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference. IMA Journal of Management Mathematics 20(2):133-145.

Koopman, SJ, Lit R. and A. Lucas. 2017. Intraday stochastic volatility in discrete price changes: the dynamic Skellam model. Journal of the American Statistical Association. 112(520):1490-503

References

Lewis, J. W., P. E. Brown, and M. Tsagris. 2017. Skellam: Densities and Sampling for the Skellam Distribution version 0.2.1 from R-Forge. https://rdrr.io/rforge/skellam/. Retrieved January 12, 2023.

Liu, X., and K. Pelechrinis. 2021. GitHub - xinliupitt/skellam regression. https://github.com/xinliupitt/skellam regression. Retrieved January 12, 2023. . 2021b. Excess demand prediction for bike sharing systems. PLoS ONE 16(6): e0252894. https://doi.org/10.1371/journal.pone.0252894.

Moreau, J.-P. 2011. Program to Calculate the First Kind Modified Bessel Function of Integer Order N, for Any REAL X, Using the Function BESSI(N,X). http://jean-pierre.moreau.pagesperso-orange.fr/Cplus/tbessi cpp.txt. Retrieved January 12, 2023.

Skellam, J. G. 1946. The frequency distribution of the difference between two Poisson variates belonging to different populations. Journal of the Royal Statistical Society (Series A) 109(3): 296.