Skellam regression in Stata

Vincenzo Verardi
LouRIM, UCLouvain
joint with Catherine Vermandele

Stata Belgian Users Meeting
(Brussels - KULeuven)

September 9, 2025

Introduction

The **Skellam distribution** models the **difference** between two **Poisson variables** with possibly **different means**. It remains valid even if the variables share an additive component, as this cancels out in the difference.

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

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

Several disciplines, including **astronomy**, **business**, and **sports**, use it to represent the difference between two counts

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 presentation we show how to **write the ML problem** and get the **gradient** and **Hessian** for numerical optimization. We then present a **Stata command** we coded.

Some examples in the literature

The Skellam distribution is often used to model the **number of points that separate two teams in sports** such as hockey and soccer.

Kendall (1951) and Dobbie (1961) show that it can also 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 neither taxis nor customers are waiting**, or **negative** if **customers are waiting**.

More recently, Liu and Pelechrinis (2021) look at the case of **shared transportation** (cars, bikes, etc.). 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 t $I_k(x)$. We are only interested in the case where $k \in \mathbb{Z}$ 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 $\left(\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt\right)$.

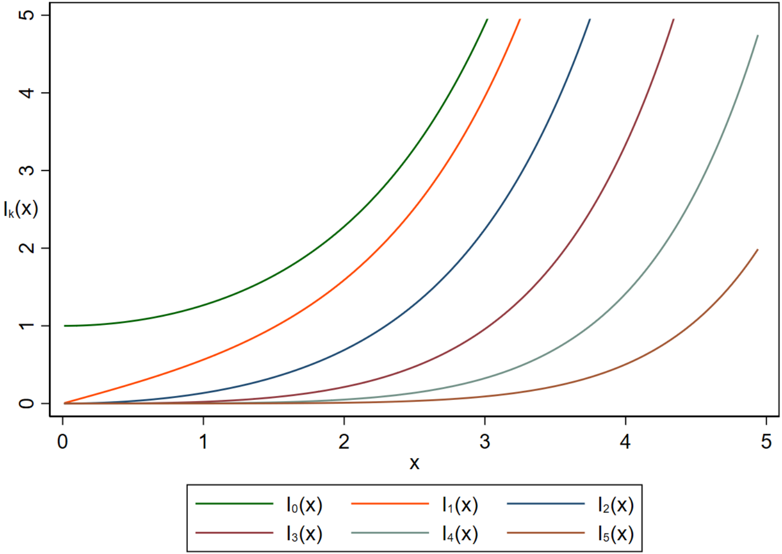
If the **values of** k **are integers** (as in our setup), $I_{-k}(x) = I_k(x)$ (see Abramowitz and Stegun 1972, p. 375, 9.6.6). $I_k(\cdot)$ can thus be replaced by $I_{|k|}(\cdot)$ in the above formula.

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** but we translated (with permission) the C++ code by Moreau (2011) to Mata (the syntax is almost identical).

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

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

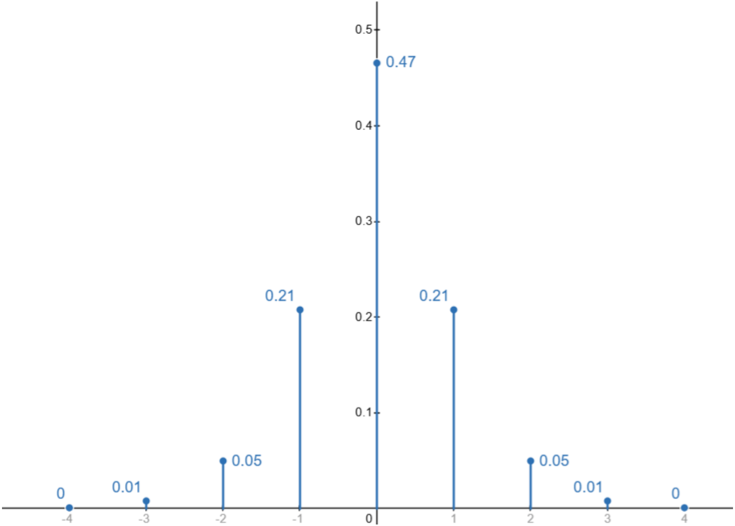
where $k \in \mathbb{Z}$ and where $I_k(\cdot)$ is the modified Bessel function of the first kind.

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

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

The mean is μ_1-μ_2, the variance is $\mu_1+\mu_2$, skewness is $\frac{\mu_1-\mu_2}{(\mu_1+\mu_2)^{3/2}}$ and kurtosis is $3+\frac{1}{\mu_1+\mu_2}$

Skellam distribution



Skellam, Bessel

Maximum Likelihood estimator

The likelihood function is given by

$$\mathscr{L}(\lambda_1, \lambda_2; k_1, \ldots, k_n) = \prod_{i=1}^{n} \Pr(Y_i = k_i \mid \lambda_1, \lambda_2)$$
$$= \prod_{i=1}^{n} \left\{ e^{-\left(e^{\lambda_1} + e^{\lambda_2}\right)} \left(e^{\lambda_1 - \lambda_2}\right)^{\frac{k_i}{2}} I_{|k_i|}\left(2\sqrt{e^{\lambda_1 + \lambda_2}}\right) \right\}$$

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

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

where

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

Maximum Likelihood estimator

To solve this maximization 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 easily 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 maximization 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 above 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: Verardi and Vermandele (2024)

```
skellamreg depvar (indepvars1) (indepvars2) [if] [in] , [options]
```

options	Description
robust	compute robust 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 constant.[1]
stub(*string*)	provide a stub for the dependent variable.[2]
technique(*string*)	change optimization technique. See [M-5] optimize##i_technique.[3]
nodofcorrection	do not correct for the degrees of freedom
level(*cilevel*)	set the confidence level

`robust` and `cluster()` options should be used with caution as the model is non-linear

Note: If only one set of explanatory variables is declared without parentheses, explanatory variables are assumed to be the same for the two underlying Poisson equations. If no explanatory variable is declared, only a constant is considered among regressors (which brings to the unconditional estimation of rate parameters). See help file for further explanations.

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 I

```
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

skellamreg 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 I

```
. skellamreg 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

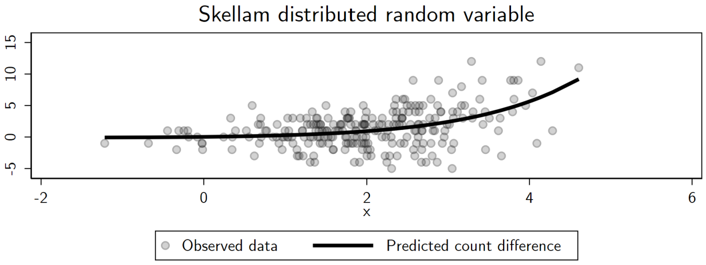
```
. test [y_count_1]:x=[y_count_2]:x=0

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

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

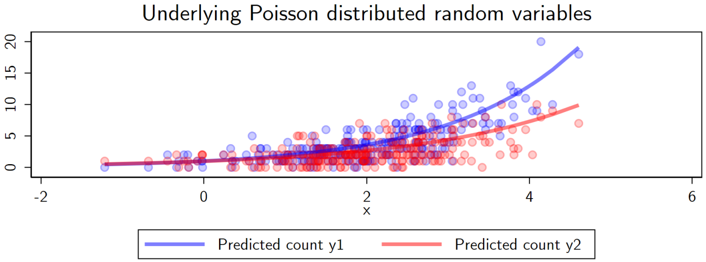
Synthetic data example I

```
. 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 I





Synthetic data example II

```
set obs 1000
set seed 1234
gen T=uniform()>0.5
gen x=rnormal(2,1)
gen y1=rpoisson(exp(0.5*x))
gen y2=rpoisson(exp(0.5*x+T))
gen dy=y1-y2
skellamreg dy x T
margins, at(T=0) at(T=1) contrast(at(r) nowald)
margins, dydx(x) predict(n1) at(T=0)
margins, dydx(x) predict(n2) at(T=0)
margins, dydx(x) at(T=0)
```

Synthetic data example II

```
. skellamreg dy x T
Iteration 0:   f(p) = -5873.3404
Iteration 1:   f(p) = -2662.7557
Iteration 2:   f(p) = -2432.5048
Iteration 3:   f(p) = -2431.8021
Iteration 4:   f(p) = -2406.7405
Iteration 5:   f(p) = -2403.3822
Iteration 6:   f(p) = -2403.1277
Iteration 7:   f(p) = -2403.1273
Iteration 8:   f(p) = -2403.1273
                                                Number of obs =     1000
```

dy	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
dy_count_1						
x	.4836783	.0447879	10.80	0.000	.3958956	.571461
T	-.2669861	.1531178	-1.74	0.081	-.5670915	.0331192
_cons	.0809609	.1185077	0.68	0.494	-.1513099	.3132316
dy_count_2						
x	.4975361	.0235597	21.12	0.000	.4513599	.5437122
T	.9563405	.0815153	11.73	0.000	.7965735	1.116107
_cons	.010138	.083723	0.12	0.904	-.1539561	.1742321

Synthetic data example II

```
. margins, at(T=0) at(T=1) contrast(at(r) nowald)

Contrasts of predictive margins                    Number of obs = 1,000

Expression: predict()
1._at: T = 0
2._at: T = 1

------------------------------------------------------------
             |            Delta-method
             |   Contrast   std. err.     [95% conf. interval]
-------------+----------------------------------------------
         _at |
   (2 vs 1)  |    -5.6799   .1838018     -6.040145   -5.319655
------------------------------------------------------------
```

Synthetic data example II

```
. margins, dydx(x) predict(n1) at(T=0)

Average marginal effects                         Number of obs = 1,000

Expression: predict(n1)
dy/dx wrt:  x
At: T = 0
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
x	1.544134	.1918626	8.05	0.000	1.16809	1.920178

```
. margins, dydx(x) predict(n2) at(T=0)

Average marginal effects                         Number of obs = 1,000

Expression: predict(n2)
dy/dx wrt:  x
At: T = 0
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
x	1.531544	.1380188	11.10	0.000	1.261032	1.802056

Synthetic data example II

```
. margins, dydx(x) at(T=0)

Average marginal effects                                   Number of obs = 1,000

Expression: predict()
dy/dx wrt:  x
At: T = 0
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
x	.0125902	.1151316	0.11	0.913	-.2130636	.238244

Football example

This case study examines how the **day of the week** a match is played may affect the **dynamics of goal scoring** in association football (soccer).

We use information from the **English Football Premier League**'s from the **season 2007-2008 to** the season **2021-2022**. All data come from https://www.footballdata.co.uk.

The dependent variable (**dftg**) that we calculate corresponds, for each match, to the difference between the number of **goals scored** by the **home** team (**fthg**) and the number of **goals scored** by the **visiting** team (**ftag**).

Bet365 **odds** are incorporated in the regression model and serve as an **indirect measure** for taking into account the **comparative strength** of teams in the game

See Karlis and Ntzoufras (2008) for details on goal modelling in football.

Football example

```
skellamreg dftg b365h b365d b365a i.day i.t, stub(ftg_)
```

	Goals	
	Home	Away
B365H	-0.206***	0.0168
	(-10.13)	(0.82)
B365D	0.0861***	0.0763
	(2.74)	(1.33)
B365A	-0.00444	-0.129***
	(-0.42)	(-6.02)
Monday	-0.0645	-0.166*
	(-0.87)	(-1.68)
Tuesday	-0.0372	-0.0926
	(-0.50)	(-0.98)
Wednesday	-0.129**	-0.0788
	(-2.07)	(-1.02)
Thursday	-0.248*	-0.0939
	(-1.85)	(-0.62)
Friday	-0.0748	0.116
	(-0.53)	(0.80)
Saturday	-0.0839**	0.0113
	(-2.17)	(0.24)
Observations	6080	
Pseudo-R^2	0.0702	
Log-likelihood	-11491.6	

t statistics in parentheses

Time dummies coefficients not reported

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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.

Bulla J, Chesneau C. and M. Kachour. 2015. On the bivariate Skellam distribution. Communication in Statistics Theory and Methods 44(21):4552–67

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 N. M. Khan. 2023. Models for Integer Data. Annual Review of Statistics and Its Application 10: 297-323.

References

Karlis D. and I. Ntzoufras. 2006. Bayesian analysis of the differences of count data. Statistics in Medicine 25(11):1885–905.

Karlis D. and I. Ntzoufras. 2008. Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. Ima Journal of Management Mathematics 20: 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

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.

Lishamol T. and G. Veena. 2022. A Retrospective Study on Skellam and Related Distributions. Austrian Journal of Statistics, 51(1), 102–111.

References

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.

Shahtahmassebi G. and R. Moyeed. 2016. An application of the generalized Poisson difference distribution to the Bayesian modelling of football scores. Statistica Neerlandica 70(3):260–73.

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.

Verardi, V. and C. Vermandele. 2024. Estimating Skellam distribution and regression parameters in Stata. The Stata Journal, 24(2): 287-300.