Late-Night TIRF Time Series with All Drinking Classes

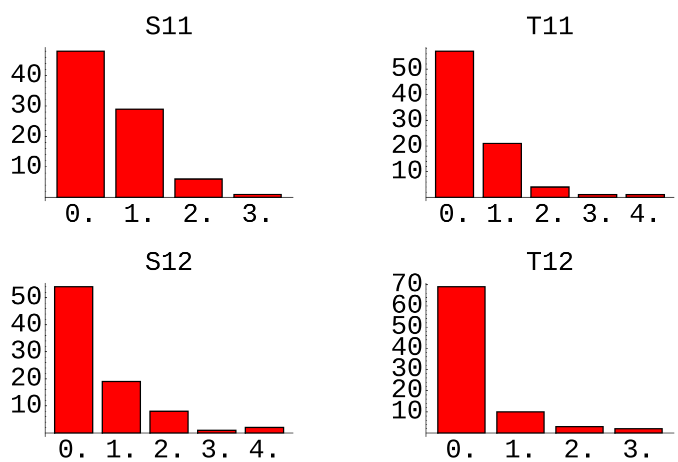
Introduction

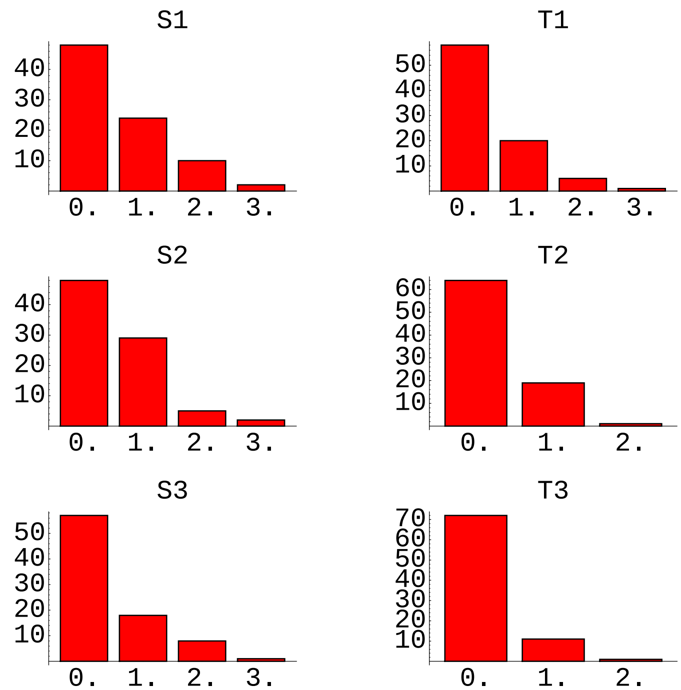
TIRF, Traffic Injury Research Foundation, provided this data on fatal car accidents in Ontario from January 1,1992 to December 31, 1998. This data are for automobile driver deaths only. The data for drinking classes "yes", "no" and "unknown" are combined in this analysis.

Ten time series were created from the TIRF dataset corresponding the two weekgroup variables SunWed and ThuSat and the five hour one hour periods beginning at 11PM, 12AM, 1AM, 2AM and 3AM. For brevity we will refer to these time series using the codes S11, S12, S1, S2, S3, T11, T12, T1, T2 and T3. The time series were aggregated to a monthly level starting January 1992 and running to December 1998. There are $n = 84$ consecutive observations in total for each time series. In our analysis of these latenight time series we are primarily interested in testing to see if a change occurred starting effective with May 1996, the 53rd observation.

Bar Chart Summaries

The TIRF late-night time series are comprised of small numbers mostly zeros. The Bar Charts look very similar to data from a Poisson distribution. However the data for T11, S12 and T12 are over-dispersed as is confirmed by the Poisson dispersion test.





Autocorrelation Analysis

The sample autocorrelation at lag k is defined by,

$$r_k = \frac{\sum_{t=k+1}^{n} (z_t - \bar{z})(z_{t-k} - \bar{z})}{\sum_{t=k+1}^{n} (z_t - \bar{z})^2}, \; k = 0, \; 1, \; 2, \; \ldots$$

provides a fairly robust test for possible serial dependence even for data which is highly discrete such as the TIRF late-night time series. If there is possible dependence we would expect it to be strongest at lag one or possibly at the seasonal lag of 12.

The standard deviation of the lag one autocorrelation coefficient is $1 \big/ \sqrt{84}$ and the benchmark significance limits are $1.96 \big/ \sqrt{84} \doteq 0.213$.

The table below gives r_1 and we see that there is no evidence of an autocorrelation in these time series at lag one.

S11	0.0430784
S12	0.127372
S1	0.181458
S2	−0.0325027
S3	0.0587406
T11	−0.0990359
T12	−0.0242566
T1	−0.0032918
T2	−0.00352113
T3	0.0210707

The table below shows r_{12}, only SunWed-2AM window shows significant seasonal correlation at the 5% level.

S11	0.00960531
S12	−0.0303431
S1	−0.0431046
S2	**0.255263**
S3	−0.0127046
T11	0.0508326
T12	−0.0449036
T1	0.0368116
T2	−0.0985915
T3	0.0127013

In view of these results, we may assume that time series are approximately statistically independent.

Poisson Modelling

Poisson Dispersion Test

We test if the pre-intervention data (ie. the first 52 observations) are approximately Poission distributed. Let z_t, $t = 1, \ldots, 52$ denote the values in the series. Then the Poisson dispersion test is based on the statistic,

$$d = \frac{\sum_{i=1}^{n} (z_t - \bar{z})^2}{\bar{z}}$$

where $\bar{z}$ is the sample mean and $n = 52$. Under the null hypothesis that the data are independent Poisson random variables, d, is distributed approximately as χ^2 on $n - 1$ df. The table below suggests that the data for the SunWed group are Poisson but there is a strong indications over over-dispersion in **S12, T11** and **T12**.

	d	p-value
S11	38.	0.911302
S12	**76.6667**	**0.0115418**
S1	45.8	0.679696
S2	48.1034	0.589411
S3	54.8947	0.329282
T11	**71.2222**	**0.0321581**
T12	**78.8182**	**0.00747674**

T1	56.12	0.288949
T2	42.5	0.795739
T3	47.	0.633224

■ Poisson Model

We will use the notation $z_t \sim \mathrm{IPo}(\lambda_t),\ t = 1,\ \ldots,\ n$ to mean that the random variables $z_t,\ t = 1,\ \ldots,\ n$ are independently distributed Poisson random variables with means λ_t. Then our intervention analysis model may be written, $z_t \sim \mathrm{IPo}(\lambda_t),\ t = 1,\ \ldots,\ n$, where $n = 84$ and

$$\lambda_t = \begin{cases} \lambda, & t = 1,\ \ldots,\ 52 \\ \lambda + \delta, & t = 53,\ \ldots,\ 84 \end{cases}$$

The null hypothesis of no effect is then $\mathrm{H}_0 : \delta = 0$. The exact log likelihood function for our model may be written as

$$\mathcal{L}(\lambda,\ \delta) = \sum_{t=1}^{n} z_t \log(\lambda_t) - \sum_{t=1}^{n} \lambda_t$$

This function may be maximized numerically to obtain the maximum likelihood estimates for λ and δ, which may be denoted by $\hat{\lambda}$ and $\hat{\delta}$. Then the null hypothesis $\mathrm{H}_0 : \delta = 0$ may be tested using a likelihood ratio test and a confidence interval for the parameter δ may be given. Under the null hypothesis $\mathrm{H}_0 : \delta = 0$ the loglikelihood function simplies to

$$\mathcal{L}(\lambda,\ 0) = \sum_{t=1}^{n} z_t \log(\lambda) - n\,\lambda$$

which is maximized with $\hat{\lambda}_0 = \bar{z}$ where $\bar{z}$ denotes the sample mean. The likelihood ratio statistic may be written,

$$R = 2\,(\max_{\lambda,\delta}\ \mathcal{L}(\lambda,\ \delta) - \max_{\lambda}\ \mathcal{L}(\lambda,\ 0)).$$

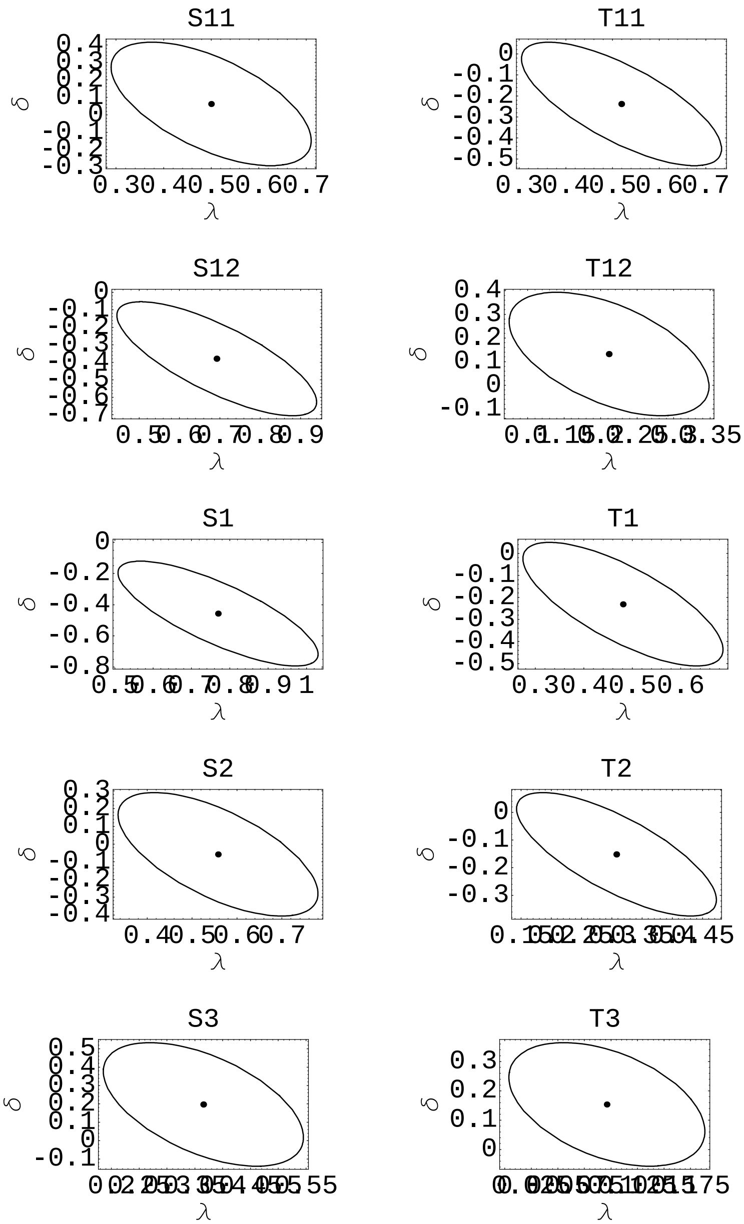
Under $\mathrm{H}_0 : \delta = 0$, R is χ^2- distributed on 1 df. From the table below we see that there is evidence that $\delta \neq 0$ for the S12, S1, T1 and T2 windows at the 10% level. In all these cases $\delta < 0$.

■ Fitted Parameters, Standard Errors, R and p-value

	λ	seλ	δ	seδ	R	p-value
S11	0.5	0.098	0.062	0.165	0.146	0.702
S12	**0.692**	**0.115**	**-0.38**	**0.152**	**5.661**	**0.017**
S1	**0.769**	**0.122**	**-0.457**	**0.157**	**7.627**	**0.006**
S2	0.558	0.104	-0.058	0.162	0.124	0.725
S3	**0.365**	**0.084**	**0.197**	**0.157**	**1.701**	**0.192**
T11	**0.519**	**0.1**	**-0.238**	**0.137**	**2.78**	**0.095**
T12	0.212	0.064	0.132	0.122	1.284	0.257
T1	**0.481**	**0.096**	**-0.231**	**0.131**	**2.865**	**0.091**
T2	**0.308**	**0.077**	**-0.151**	**0.104**	**1.944**	**0.163**
T3	**0.096**	**0.043**	**0.154**	**0.098**	**2.914**	**0.088**

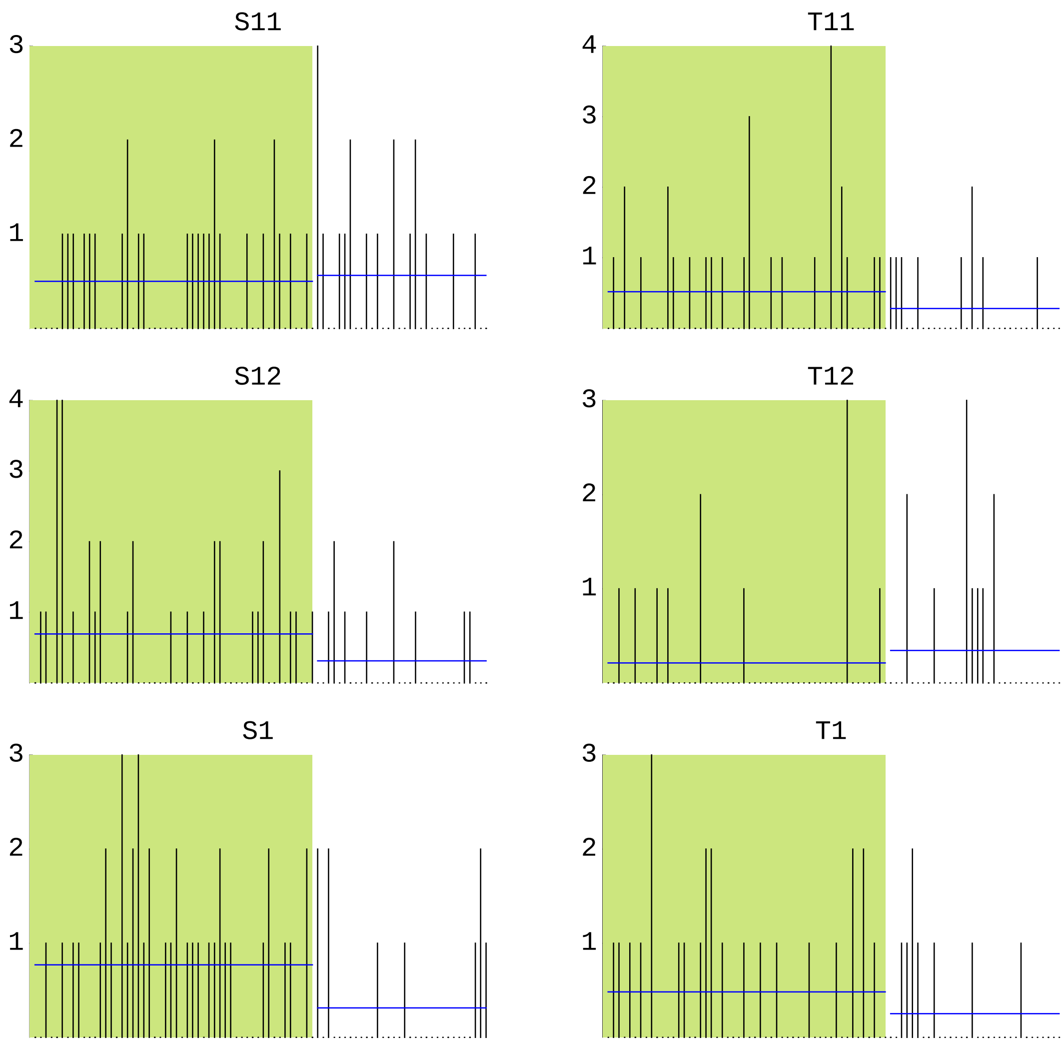
The p-value reported is for a two-sided test. A one-sided test would seem to be more appropriate so in this case the p-value in the above table should be halved.

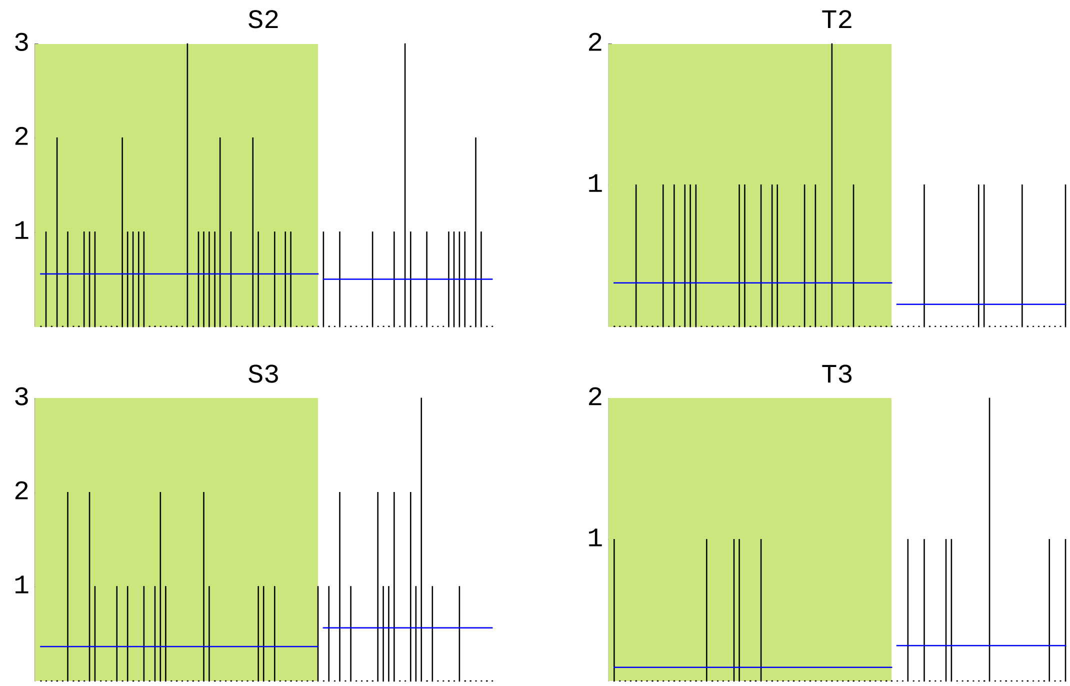
■ 90% Confidence Ellipses for λ and δ



■ Fitted Values and Visualization

The expected values of z_t in our model is given $\mathbb{E}\{z_t\} = \lambda_t = \lambda + \delta\,\xi_t$. The expected value is shown as a blue line in the graphs below.





Negative Binomial Modelling

■ Negative Binomial Distribution

Actual count data are often over-dispersed, that is, they fail the Poissiion dispersion test. In this case, the negative binomial distribution provides a more flexible alternative than the Poisson distribution for modelling discrete random variables. Suppose there is an unobserved random variable E having a gamma distribution with mean 1 and variance $1/\theta$ and that conditional on E, the random variable Y has a Poisson distribution with mean μE. Then Y has a negative binomial distribution and its density function may be written,

$$f(y;\theta,\mu) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)\,y!}\;\frac{\mu^y\,\theta^\theta}{(\mu+\theta)^{\theta+y}}$$

The mean and variance of Y are given by $\mathcal{E}\{Y\} = \mu$ and $\mathrm{Var}\{Y\} = \mu + \mu^2/\theta$. Notationally we may denote this distribution by NB(μ, θ).

■ Generalized Linear Models

The generalized linear model provides an alternative and more general statistical model for these data. GLM's are frequently used for regression modelling of non-Gaussian data such as data arising from the binomial, lognormal or negative binomial distributions. Given independently distributed z_t, $t = 1, \ldots, n$ and possibly p covariates of interest $x_{t,j}$, $t = 1, \ldots, n$, $j = 1, \ldots, p$ the GLM may be defined. There are three components to a GLM:

(i) the statistical density or probability function, $f(z_t;\ \mu_t;\ \theta)$, where θ denotes distributional parameters, $\mu_t = \in\{z_t\}$ and it is assumed that μ_t depends on the distribution parameter or parameters as well as the covariates.

(ii) the linear predictor which depends on the covariates linearly,

$$\eta_t = \sum_{j=1}^{p} \alpha_t \, x_{t,j}$$

(iii) the link function, $\eta_t = \ell(\mu_t)$.

The standard GLM algorithm is based on Iteratively Reweighted Least Squares (IRLS) and this algorithm provides a good approximation to the more exact maximum likelihood method. Using *Mathematica* it is possible to obtain the exact maximum likelihood estimates which are preferable to the IRLS estimates.

■ Model Formulation

For $j = 1$, we take $x_{t,1} = 1$, $t = 1, \ldots, n$ which corresponds to the overall mean. The intervention is represented by,

$$x_{t,2} = \xi_t = \begin{cases} 0 & t \leq 48 \\ 1 & t > 49 \end{cases}$$

It is assumed that $z_t \sim \mathrm{NB}(\lambda_t,\ \theta)$ where

$$\log(\lambda_t) = \lambda + \delta\, \xi_t$$

that is, the link function is taken to be logarithmic.

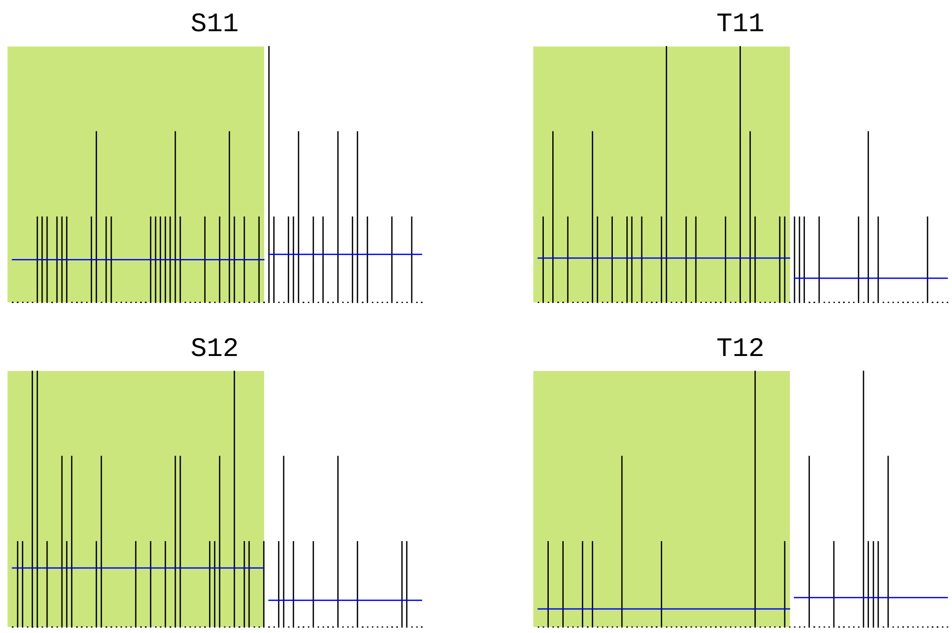
Maximum Likelihood Estimates

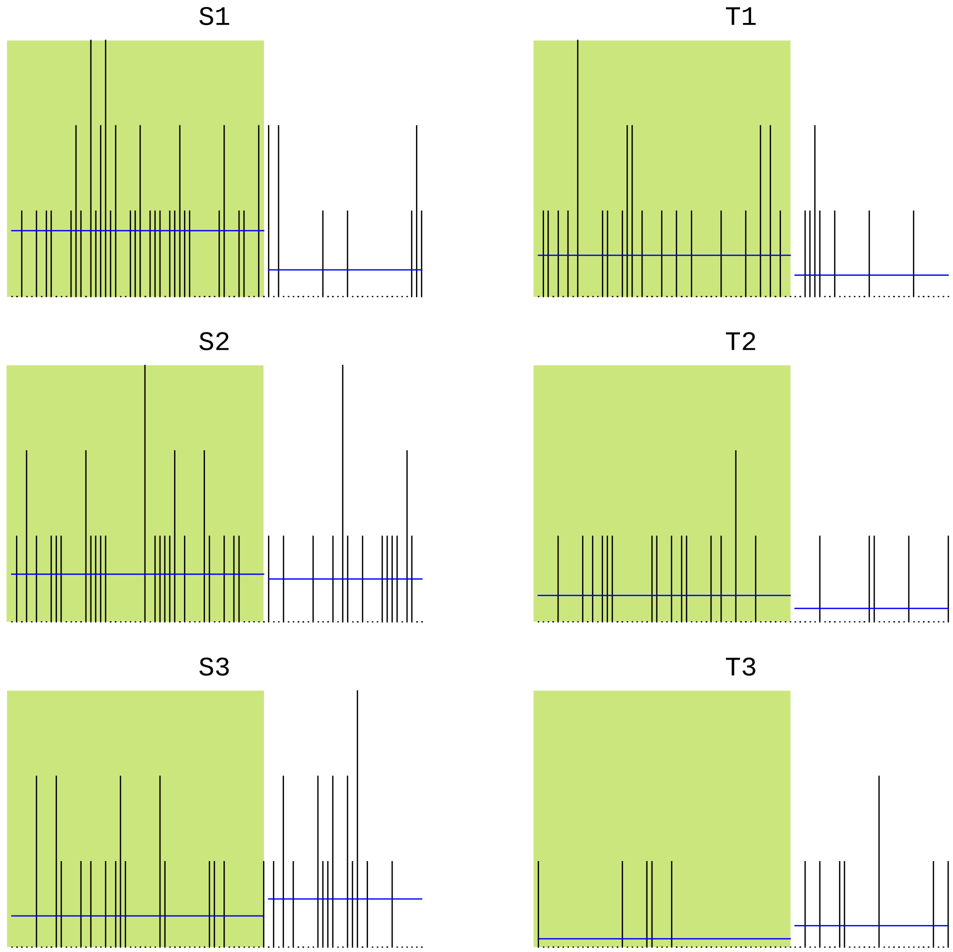
	λ	δ	θ	R	p-value
S11	-0.693	0.118	148533.	0.146	0.702
S12	**-0.368**	**-0.795**	**1.407**	**4.038**	**0.044**
S1	**-0.262**	**-0.901**	**23233.4**	**7.529**	**0.006**
S2	-0.584	-0.109	430.549	0.119	0.73
S3	-1.007	0.431	2.565	1.426	0.232
T11	-0.655	-0.613	1.953	2.269	0.132
T12	-1.553	0.485	0.399	0.757	0.384
T1	**-0.732**	**-0.654**	**6.288**	**2.661**	**0.103**
T2	**-1.179**	**-0.678**	**313336.**	**1.944**	**0.163**
T3	**-2.338**	**0.951**	**1054.07**	**2.913**	**0.088**

The p-value reported is for a two-sided test. A one-sided test would seem to be more appropriate so in this case the p-value in the above table should be halved.

Visualization

The expected values of z_t in the negative binomial regression model is given by $\mathbb{E}\{z_t\} = e^{\lambda+\delta\,\xi_t}$. As expected the impact of the interventions is almost the same as that given the by Poisson model.





Comparison with Normal Regression Modelling

The late night time series are comprised of very small numbers and these number clearly violate the assumption of normality as the bar charts made clear. However normal regression model would be expected to be fairly robust against such departures as shown by Hjort (1994). It is of interest to compare our previous analyses using Poisson and Negative Binomial regression with standard normal regression. In the standard normal regression we may formulate our step intervention model,

$$z_t = \mu + \delta\, \xi_t + N_t \tag{1}$$

where N_t is the error term. Based on the pre-intervention data we assume initially that N_t is normal and independent, so ordinary multiple linear regression can be used. The intervention series are defined by,

$$\xi_t = \begin{cases} 0 & t < 53 \\ 1 & t \geq 53 \end{cases}$$

The term N_t represents the disturbance or error term and it has been tentatively identified as Gaussian white noise, that is $N_t = a_t$, where $a_t \sim \text{NID}(0,\ \sigma^2)$.

The following table is in quite close agreement with the results from the Poisson analyses. However only 6 interventions are detected on a one-sided test at the 10% level whereas previously there were 7 and significance levels are larger suggesting this analysis is not quite as sensitive as the Poisson analysis. There is almost no difference though in many cases such as for T2 and T3.

		Estimate	SE	TStat	PValue
S11	1	0.5	0.0954831	5.23653	1.23913×10^{-6}
	ξ	0.0625	0.1547	0.404007	0.687259
S12	1	0.692308	0.122467	5.65303	2.22552×10^{-7}
	ξ	**-0.379808**	**0.198419**	**-1.91417**	**0.0590865**
S1	1	0.769231	0.106216	7.24214	2.16742×10^{-10}
	ξ	**-0.456731**	**0.17209**	**-2.65403**	**0.00955079**
S2	1	0.557692	0.100219	5.56474	3.21682×10^{-7}
	ξ	-0.0576923	0.162373	-0.355307	0.723272
S3	1	0.365385	0.0991671	3.68453	0.000409787
	ξ	0.197115	0.160669	1.22684	0.223393
T11	1	0.519231	0.103242	5.02926	2.85058×10^{-6}
	ξ	**-0.237981**	**0.167271**	**-1.42273**	**0.158609**
T12	1	0.211538	0.0891536	2.37274	0.0199973
	ξ	0.132212	0.144445	0.915305	0.362715
T1	1	0.480769	0.0905745	5.308	9.26536×10^{-7}
	ξ	**-0.230769**	**0.146748**	**-1.57256**	**0.119672**
T2	1	0.307692	0.0636884	4.83122	6.22699×10^{-6}
	ξ	**-0.151442**	**0.103187**	**-1.46765**	**0.146025**
T3	1	0.0961538	0.0541851	1.77454	0.0796849
	ξ	**0.153846**	**0.08779**	**1.75243**	**0.0834372**

References

Hjort, N.L. (1994), The exact amount of t-ness that the normal model can tolerate, *Journal of the American Statistical Association* 89, 665-675.