ROBUST FITTING FOR THE NUMBER OF DEFECTS AS A MEASURE OF PRODUCT QUALITY

HANAN ELSAIED

Department of Statistics, Faculty of Commerce, Suez Canal University,
Ismailia

Abstract

The number of defects, as a count time series data, is an important measure of product quality which is widely used in industry. We discuss M-estimation of INARCH-models as appropriate models for analysis and modeling such data, especially, in the presence of outliers. These models are proposed by Elsaied and Fried (2014) for conditional Poisson distributions. They compare between the performance of the conditional maximum likelihood estimator (CML) and Tukey M-estimators with and without bias correction. Liboschik et al. (2017) construct the tscount package in the R programming language, which provide likelihood-based estimation methods for analysis and modeling of count time series based on generalized linear models. In our paper, we compare between the results of the best functions, which are built in the R programming language, for the Tukey M-estimators in the case of the Poisson INARCH(1) model given in Elsaied and Fried (2014) and the tsglm function in the tscount package. We investigate the performance of these estimators under assuming different outliers scenarios by simulations. The usefulness of the chosen functions is applied on a real defects data example.

Keywords: Defects count data; Poisson model; INARCH models; GLM models; Tukey M-estimator; Additive outliers; Robustness.

1. Introduction

In the last two decades, count time series data have received increasing attention. They are found in many different applications, e.g. from medicine, finance or industry. Integer-valued GARCH (briefly: INGARCH) models have been studied by Ferland et al. (2006) and Fokianos et al. (2009), among others, as appropriate models for thise type of data. Fokianos and Fried (2010) model different types of outliers and interventions in INGARCH-processes and propose an iterative procedure for the detection of such effects.

Elsaied and Fried (2012) provide INGARCH(1,0) or more briefly called INARCH(1) model for count time series as a simplest version of the INGARCH models under assuming the observation at each point in time to follow a Poisson distribution conditionally on the past, with the conditional mean being a linear function of previous observations. They construct new M-estimators based on the Tukey function as modified versions of maximum likelihood estimators to fit count data robustly, where the Poisson model provides a standard framework for the analysis of this type of data.

Elsaied and Fried (2014) build some functions in R-programming language for these estimators and perform some simulation experiments to compare the performance of them. They compare the following procedures:

1. The conditional maximum likelihood estimator (CML) with initialization from an autoregressive, AR(1), fit.
2. The glmrob R-function with robust initialization from an AR(1) fit, see package robustbase.
3. The uncorrected bias Tukey M-estimator with robust initialization from an AR(1) fit and tuning constant $k \in \{5,7\}$, the tuning constant k regulates the trade-of between the robustness and the efficiency of the estimators.
4. The bias-corrected Tukey M-estimator with robust initialization from an AR(1) fit and tuning constant $k \in \{5,7\}$, called Tukeycorr in their notations.

Liboschik et al. (2017) construct the R package tscount which provides likelihood-based estimation methods for analysis and modeling of count time series based on generalized linear models. The package includes methods for model fitting, prediction and intervention analysis.

In this paper, we compare the performance of the CML, the best function (Tukeycorr), which is given above in Elsaied and Fried (2014), for the Tukey M-estimators in the case of the Poisson INARCH(1) model and the tsglm function in the tscount package, which is given in Liboschik et al. (2017) via simulations in case of outlier-contaminated Poisson time series data.

Section 2 defines the defects count data. We model and estimate the defects count data in Section 3. We compare the performance of the chosen functions via simulations in Section 4. We apply these functions on a real defects data example in Section 5. Section 6 gives some conclusions.

2. Defects count data

There are many definitions for the product quality. One definition of the product quality relates it to decrease the number of the defects product. When there are no defects, it means what we call "zero defects" concept. In a practical sense, this concept is hard to reach. Alternatively, the scientifics suggest applying the six sigma concept to decrease the number of the defects product. Six Sigma stands for 6 standard deviations (6s) between the average and the acceptable limits of the quality levels, for more details, see Breyfogle (1999) and Bertolaccini et al. (2015). Either we agree to this concept or not, we should work to decrease the number of the defects product as a measure of its quality. The number of the defects product, as a count variable, is one of the variables which we can statistically control by we so called “control chart, for more details, see Thomas and Lonnie (1986). Control charts are simple, robust tools for understanding the production process variability. They depend on taking random samples from a population of the products units and knowing the number of the defects product.

This paper aims to provide a statistical model for the variable of the number of defects product and estimate its parameters either its data contains outliers or not, see Figure 1.

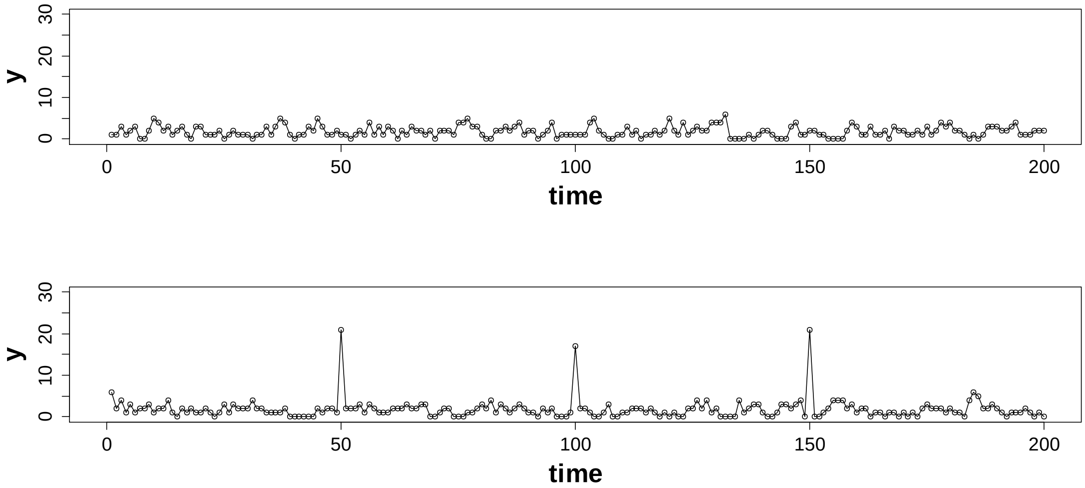

Figure 1. Simulated data from an INARCH(1) model with parameters $\boldsymbol{\beta_0 = 1}$ and $\boldsymbol{\alpha_1 = 0.4}$ in case of clean data (top) and three additive outliers (bottom).

Figure 1 (top) shows simulated clean data from an INARCH(1) model with parameters $\beta_0 = 1$ and $\alpha_1 = 0.4$ and it shows (bottom) data with three additive outliers of increasing random size generated from Poisson distributions with means varying from 1 to 20 at times 50, 100 and 150.

3. Modeling and Estimation

Let (Y_t) denotes the variable of the number defects product with range $\mathbb{N}_0$= {0, 1, ...}. The adequate distribution for Y is the Poisson distibution with mean λ and probability density function $f_\lambda(y) = \frac{e^{-\lambda}\lambda^y}{y!}$, $y \in \mathbb{N}_0$. Ferland et al. (2006) define an INGARCH (p, q) process $(Y_t\colon t \geq 1)$ of orders p and q through the relationships

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t),$$
$$\lambda_t = \beta_0 + \sum_{i=1}^{p} \alpha_i \, Y_{t-i} + \sum_{j=1}^{q} \beta_j \, \lambda_{t-j}, \qquad (1)$$

for $t \geq 1$. The dynamics of the process are modeled via the conditional mean $\lambda_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$ of Y_t, where $\mathcal{F}_{t-1}$ stands for the σ–field generated by $\{Y_{1-q}, \dots, Y_{t-1}, \lambda_{1-p}, \dots, \lambda_0\}$ representing the whole information up to time $t - 1$. Here, $\beta_0 > 0$ is an intercept and α_i, $i = 1, \dots, p$, and β_j, $j = 1, \dots, q$, are non-negative regression parameters. A stationary process fulfilling (1) with mean $\beta_0/(1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j)$ exists if $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$.

Elsaied and Fried (2012) provide INARCH(1) model as a simplest version for these models under assuming the observation at each point in time to follow a Poisson distribution conditionally on the past, with the conditional mean being a linear function of previous observations is as follows:

$$\lambda_t = \beta_0 + \alpha_1 Y_{t-1}. \tag{2}$$

The properties of INARCH(1) model have been given in Wiss (2010) as follows:

1. An INARCH(1) process with $\beta_0 > 0$ and $0 < \alpha_1 < 1$ is a stationary Markov chain with the transition probabilities
   $P_{ji} := p(Y_t = i | Y_{t-1} = j) = exp(-\beta_0 - \alpha_1 \cdot j) \cdot \frac{(\beta_0 + \alpha_1 \cdot j)^i}{i!} > 0.$
   It is irreducible and aperiodic and hence ergodic.
2. The marginal mean is $\lambda = E(Y_t) = \frac{\beta_0}{1-\alpha_1}$, and the marginal variance is
   $V(Y_t) = \frac{\beta_0}{(1-\alpha_1)(1-\alpha_1^2)}$.
3. The autocorrelation function $\rho_Y(h) = corr[Y_t, Y_{t-h}]$ equals α_1^h, like in the standard AR(1) model. An INARCH(1) model can be described by an AR(1) structure satisfying the equation $Y_t = \alpha_1 Y_{t-1} + \epsilon_t$ with ϵ_t being a white noise error term.

The data can be contaminated with one or more different types of outliers, transient, level shift or additive (spiky) outliers. When the data contains for example three additive outliers of increasing expected size generated from Poisson distributions with means varying from 1 to 20 at times 50, 100, and 150 then the INARCH(1) model can be defined by

$$Z_\tau = Y_\tau + X, \tag{3}$$

where Z_τ is the new contaminated observed value at a specific time τ and X is the size of the outlier such that $X \sim$ Poisson(ν) with $\nu \in \{1,2,\ldots,20\}$. Here, Y_t follows model (2) and $Z_t = Y_t$ if $t \neq \tau$ with $\tau = 50$, $\tau = 100$ and $\tau = 150$, see Figure 1.

According to Elsaied and Fried (2012), the conditional maximum likelihood (CML) approach is used to estimate the parameters of the INARCH(1) as the solution of the following set of estimation equations:

$$\begin{aligned} S_n(\theta) &= \sum_{t=p+1}^{n} (\frac{y_t}{\lambda_t} - 1) \frac{\partial \lambda_t}{\partial(\theta_i)} \\ &= \sum_{t=1}^{n} (\frac{y_t - \lambda_t}{\sqrt{\lambda_t}}) \frac{1}{\sqrt{\lambda_t}} \frac{\partial \lambda_t}{\partial(\theta_i)} = 0 \end{aligned} \tag{4}$$

where $y_1, \dots, y_n$ be a time series from an INARCH(1) process, $\theta = (\beta_0, \alpha_1)'$ denotes to the vector of its parameters and $\partial\lambda_t/\partial\theta = (1, y_{t-1})'$.

M-estimation for the INARCH(1) model with asymptotic bias correction $(a_0, a_1)'$ as a robust version for Equation (4) is as follows:

$$\sum_{t=2}^{n} \psi\left(\frac{y_t - \lambda_t}{\sqrt{\lambda_t}}\right)\frac{1}{\sqrt{\lambda_t}}\left(\sigma\psi\left(\frac{y_{t-1} - \lambda}{\sigma}\right) + \lambda\right) \\ -(n-1)\binom{a_0}{a_1} = \binom{0}{0}. \tag{5}$$

where observations with large standardized residuals $(y_t - \lambda_t)/\sqrt{\lambda_t}$ are truncated using Huber's or Tukey's ψ function, and do the same with regressors y_{t-1} which are outlying w.r.t. the marginal distribution with $\lambda = \beta_0/(1-\alpha_1)$ is the marginal mean and $\sigma^2 = \beta_0/((1-\alpha_1)(1-\alpha_1^2))$ is the marginal variance. For more details and the calculation methods, review Elsaied and Fried (2012).

4. Simulations

In this section, we perform some simulations to compare the performance of the CML estimator and the best-Tukey M-estimators with bias correction given in Elsaied and Fried (2012). Because the INARCH(1) model belongs to the class of generalized linear models, we include the R-function tsglm from package tscount, which is based on the work of Liboschik et al. (2017), in our comparison using the identity link, family Poisson and the lagged variables as regressors. Altogether, we compare the following procedures:

1. CML with initialization from an AR(1) fit.
2. The bias-corrected Tukey M-estimator defined in (5) with robust initialization from an AR(1) fit and tuning constant ($k = 7$), called Tukeycorr here.
3. tsglm function in tscount package to fit a generalized linear model (GLM) for time series of counts.

These estimators are compared in terms of bias and root of the mean square error (RMSE) in finite samples from an INARCH (1) model in the following cases:

1. The case of presence twenty additive outliers of fixed small size in subsection 4.1.
2. The case of presence twenty additive outliers of fixed large size in subsection 4.2.
3. The case of presence three additive outliers of of increasing size in subsection 4.3.

4-1 Twenty additive outliers of fixed small size

Here, we compare the biases in the presence of twenty additive outliers of fixed small size equal 5 considering data from an INARCH(1) model with parameters $\beta_0 = 3$ and $\alpha_1 = 0.1$. The results are based on 200 data sets of size 100.

The results for the RMSE are dominated by the bias and thus not shown here.

In Figure 2, isolated outliers cause a strong bias effect on all compared estimators, but tsglm works better than CML and Tukeycorr if the numbers of outliers more than or equal 5. When the number of outliers is less than 5, Tukeycorr works better than the others. To be sure from these results, we suggest to perform another simulation in case of presence number of outlier less than 5 for example in case of presence three outliers, see the subsection 4.3.

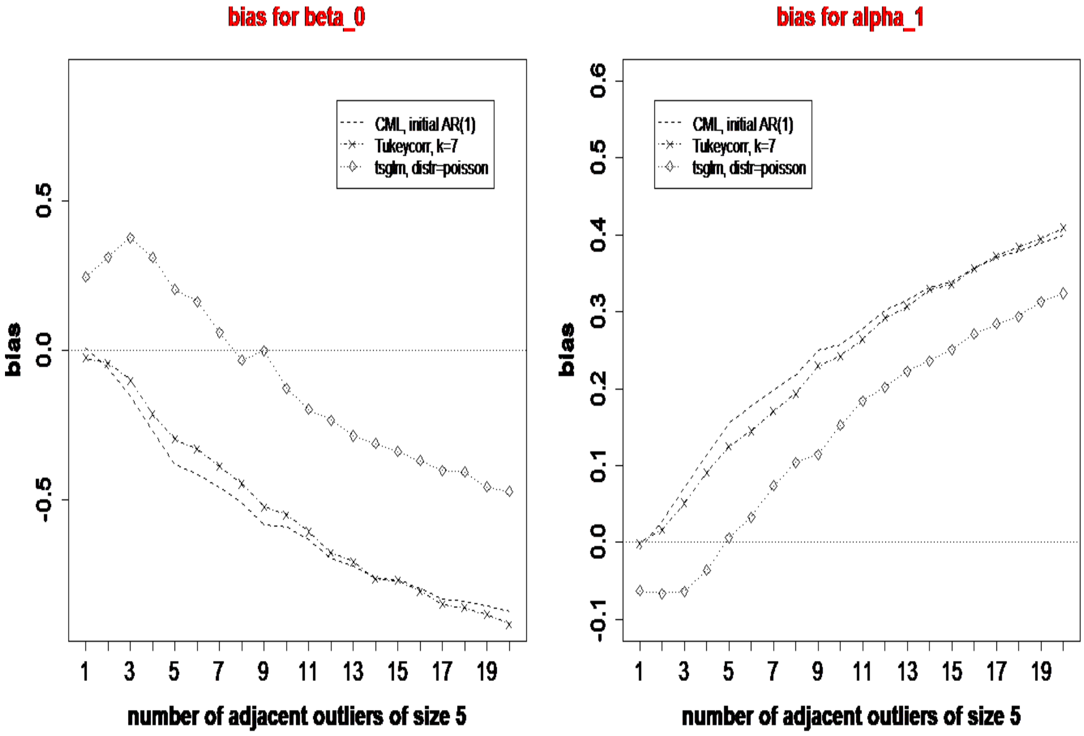

Figure 2. Simulated biases of the CML, Tukeycorr with tuning constant k = 7, and tsglm in case of twenty additive outliers of fixed small size = 5 and true values $\beta_0 = 3$ and $\alpha_1 = 0.1$, sample size n=100, for β_0 (left) and α_1 (right).

4-2 Twenty additive outliers of fixed large size

Here, we compare the biases in the presence of twenty additive outliers of fixed large size equal 15 considering data from an INARCH(1) model with parameters $\beta_0 = 3$ and $\alpha_1 = 0.1$. The results are based on 200 data sets of size 100. Also the results for the RMSE are dominated by the bias and thus not shown here.

In Figure 3, Tukeycorr gives the smallest bias if it is compared to CML or tsglm.

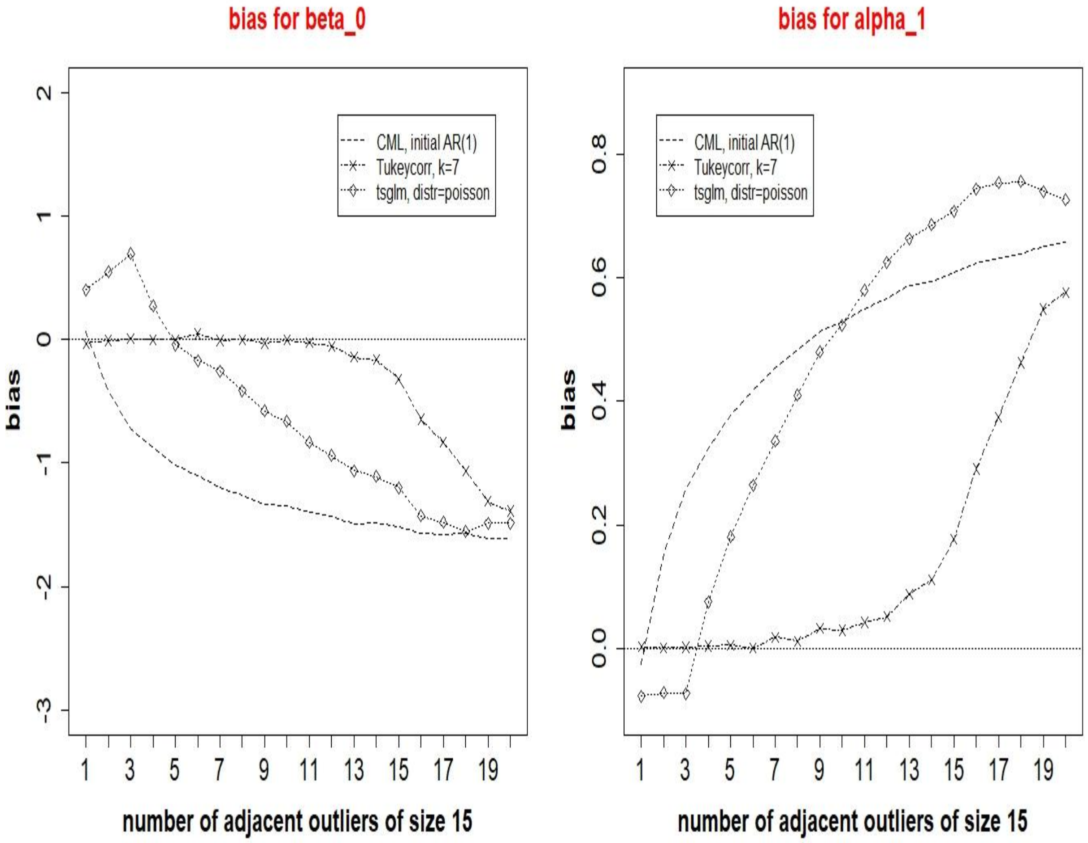

Figure 3. Simulated biases of the CML, Tukeycorr with tuning constant $\mathbf{k = 7}$, and tsglm in case of twenty additive outliers of fixed large size = 15 and true values $\boldsymbol{\beta_0 = 3}$ and $\boldsymbol{\alpha_1 = 0.1}$, sample size n=100, for $\boldsymbol{\beta_0}$ (left) and $\boldsymbol{\alpha_1}$ (right).

4-3 Three additive outliers of increasing size

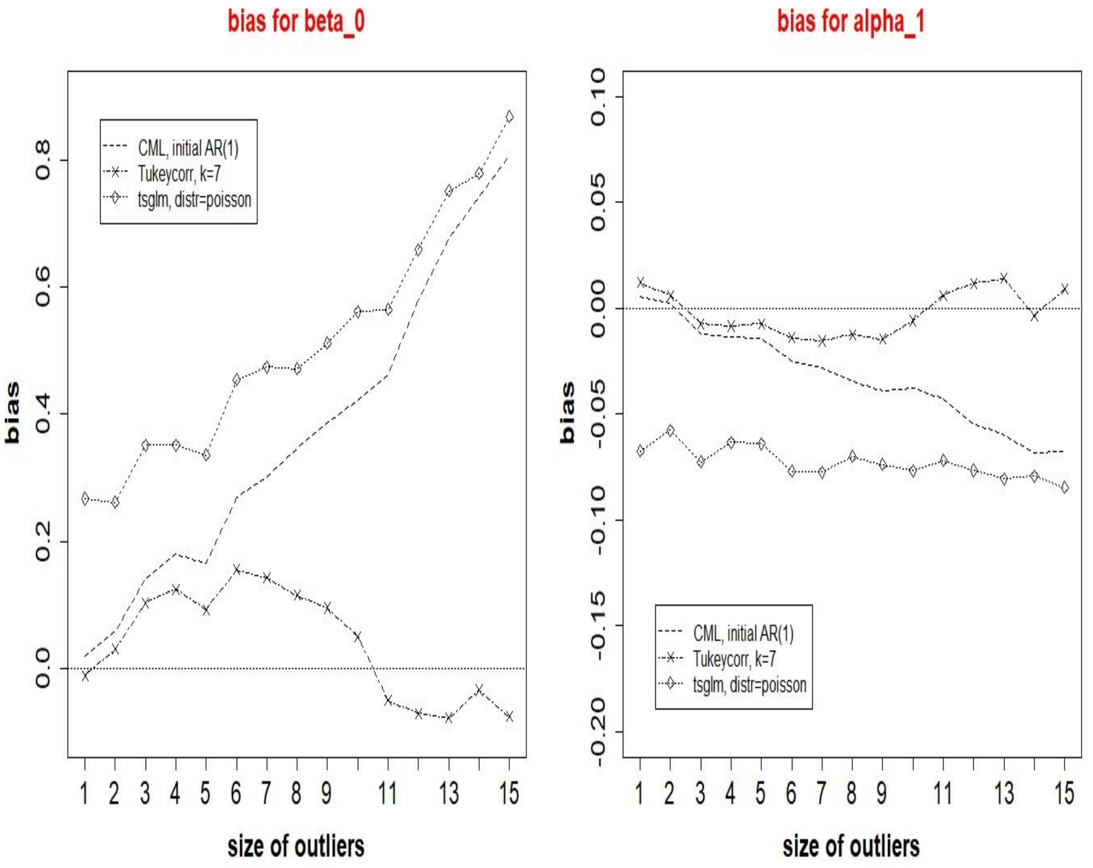

Figure 4. Simulated biases of the CML, Tukeycorr with tuning constant $\mathbf{k = 7}$, and tsglm in case of three additive outliers of increasing size and true values $\boldsymbol{\beta_0 = 3}$ and $\boldsymbol{\alpha_1 = 0.1}$, sample size n=100, for $\boldsymbol{\beta_0}$ (left) and $\boldsymbol{\alpha_1}$ (right).

Here, we compare the biases in the presence of three additive outliers at of increasing size $\{1,2,\ldots,10,12,\ldots,20\}$ at adjacent positions $\{15,30,45\}$ considering data from an INARCH(1) model with parameters $\beta_0 = 3$ and $\alpha_1 = 0.1$. The results are based on 200 data sets of size 100.

Again the results for the RMSE are dominated by the bias and thus not shown here.

In Figure 4, Tukeycorr M-estimator gives less bias than CML and tsglm for both parameters β_0 and α_1.

5. Application on a real data example

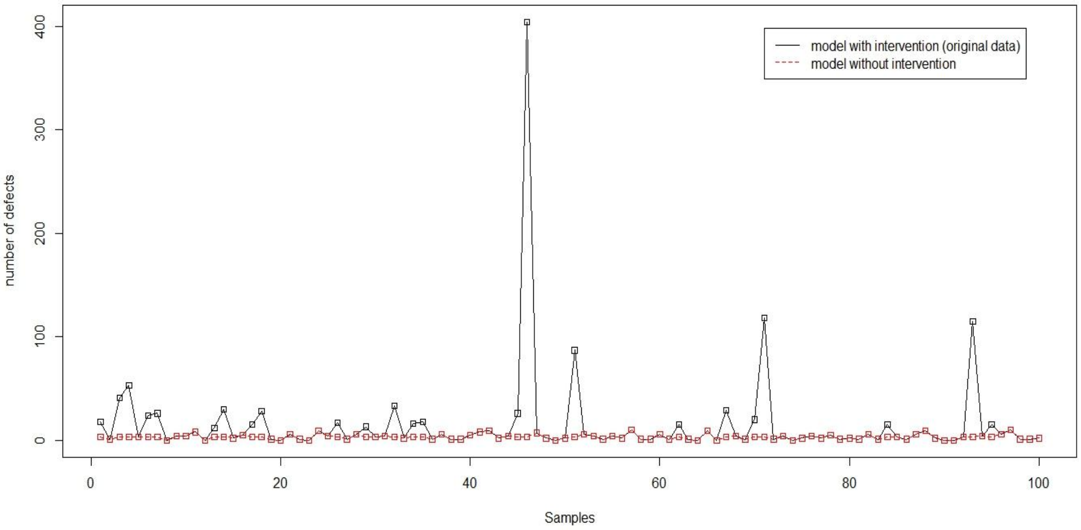

Figure 5. plywood data (solid black line) and data after removal of outliers (dotted red line).

We investigate the reliability of our proposed methods using a real data in the industry field. Manufactures laminated plastic plywood (briefly: plywood) data is given in Filho and Sant'Anna (2015). For this data, we have 100 observations for the variable of the number of defects found in produced plywoods. we observe that the data contains seven observations with values greater than 30, eighteen observations between 10 to 30 and the rest, 75 observations, less than 10. We consider that data is collected at different time points or as a count time series data, see the series of the original data in Figure 5.

To apply our proposed methods on this data, we follow the following steps:

1. We run the "interv multiple" function in tscount package to detect outliers, this function performs an iterative procedure for detection of multiple intervention effects of unknown types occuring at unknown times. The code of this function in R and its results are as follows:

```
countfit_interv <- interv_multiple(fit=tsglm(ts=y, model=list(past_obs=1)), tau
                s=1:100, deltas=c(0,0.8,1), B=500, signif_level=0.05),
```

Here, the "interv multiple" function used to detect outliers at unknown times from 1 to 100, where taus = 1:100 is an integer vector of times which are considered for the possible intervention, and for deltas = c(0,0.8,1) as a numeric vector that determines the types of intervention to be considered 0 for spiky outlier, 1 for level shift outlier and 0.8 for transient shift outlier. B is a positive integer value giving the number of bootstrap samples for estimation of the p-value and "signif level" is a numeric value between 0 and one giving a significance level for the procedure.

Detected intervention(s):

Table 1. The results of “interv multiple” function

tau	delta	size	test_statistic	p_value
52	1	1.340066e-09	150.97785	0
46	0	2.837735e+02	10215.44544	0
8	1	1.887947e-09	132.05852	0
71	0	1.086792e+02	1129.79172	0
93	0	1.068003e+02	1216.88412	0
51	0	8.004906e+01	792.67633	0
3	0.8	2.889325e+01	311.51157	0
48	1	3.234328e-10	33.69842	0
32	0	2.746446e+01	123.72798	0
2	1	1.772284e-07	26.81250	0
67	0	2.355734e+01	97.04179	0
45	0	2.078209e+01	77.85561	0
18	0	1.886202e+01	68.93868	0
14	0	1.611020e+01	51.05992	0
70	0	1.493166e+01	47.48654	0
35	0	1.338651e+01	39.97027	0
26	0	1.250249e+01	33.64359	0
34	0	1.161703e+01	30.03328	0
62	0	1.072314e+01	26.42178	0
84	0	1.081630e+01	27.67084	0
95	0	1.096644e+01	28.77609	0
29	0	9.057932e+00	20.23066	0
46	0	9.138280e+00	21.06398	0
17	0	8.148362e+00	17.42712	0

According to the results of “interv multiple” function in Table 1, we detect 24 outliers (four level shifts, one transient and 19 additives), see the series of the clean data in Figure 5

2. We fit an INARCH(1) model to observed plywoods data using conditional maximum likelihood (CML) as a non robust method, Tukey M-estimation with bias correction and tsglm as a robust estimates of the model parameters, see Table 2 (left)

3. We fit again an INARCH(1) model using the same functions, but after having cleaned the data from outliers using step 1, see Table 2 (right)

Table 2. Parameter estimates for the campy data (left) and for the cleaned campy data (right)

Estimators estimates	observed		cleaned	
	$\hat{\beta}_0$	$\hat{\alpha}_1$	$\hat{\beta}_0$	$\hat{\alpha}_1$
CML	13.3591	0.06131	2.9496	2.512065e-08
Tukeycorr	2.6518	1.877702e-10	2.7711	2.211659e-11
tsglm	14.280	2.685770e-11	2.95	8.559052e-12

From Table 2, there are small differences between CML, Tukeycorr and tsglm for both β_0 and α_1 after having cleaned the data set from outliers. Tukeycorr for the original data are closer to the estimates for the cleaned data than the CLM and tsglm estimates.

Generally, Tukey M-estimator shows better results in the presence of outliers, since the robust estimates are closer to the estimates for the cleaned data than the non-robust estimates. Also it gives better results than tsglm, since here we have nearly 25**%** outliers in the data with large sizes. This results as the same results in our simulation above, see again subsection 4.2.

There is almost no serial for the estimator of the dependence parameter α_1. An explanation of this problem is that the presence of four level shifts can be interpreted as a long sequence of outliers, i.e. a high level of contamination.

6. Conclusions

We have suggested robust INARCH(1) model as a promise model to fit the defects count data when the data contains outliers. Our simulation results indicate that in case of the INARCH(1) model, the corrected-bias Tukey M-estimator is more robust when the data contains high percent of outliers of fixed large size or when there is small number of outliers of increasing size. But in case of high percent of outliers of fixed small size tsglm works better. The application results on a real data example give similar results as our simulation, but the problem of the dependence parameter remains an open problem need to search.

Acknowledgement: We are grateful to Dr. Salah Mahdy Mohamed, who works as an assistant Professor at Cairo University, Institute of Statistical Studies and Research(ISSR), for his helping to have the paper of Filho1, D. and Sant' Anna, A. (2016), which contains the data of our paper.

References

Bertolaccini, L., Viti, A., Terzi, A. (2015) The Statistical point of view of Quality: the Lean Six Sigma methodology *Journal of Thoracic Disease. 7(4): E66– E68. doi: 10.3978/j.issn.2072-1439.2015.04.11 PMCID: PMC4419289 .*

Breyfogle, F.W. III (1999), Implementing Six Sigma: Smarter Solutions Using Statistical Methods, *John Wiley* Sons, Inc., New York, NY. [Google Scholar]

Elsaied, H. (2012). Robust Modelling of Count Data. Ph.D. thesis, Technische Universität Dortmund, Dortmund. http://hdl.handle.net/2003/29404.

Elsaied, H., Fried, R. (2014). Robust fitting of INARCH models. Journal of Time Series Analysis, 35(6), 517– 535. 10.1111/jtsa.12079.

Ferland, R., Latour, A. and Oraichi, D. (2006): Integer–valued GARCH processes. *Journal of Time Series Analysis 27, 923-942*.

Filho1, D. and Sant' Anna, A. (2016) Principal component regression-based control charts for monitoring count data. *The International Journal of Advanced Manufacturing Technology 85:1565– 1574DOI 10.1007/s00170-015-8054-6*.

Fokianos, K., Rahbek, A. and TjØstheim, D. (2009): Poisson autoregression. *Journal of the American Statistical Association 104, 1430-1439*.

Fokianos, K. and Fried, R. (2010): Interventions in INGARCH processes. *Journal of Time Series Analysis 31, 210-225*.

Fokianos, K. and Fried, R. (2010) Interventions in INGARCH processes. Journal of Time Series Analysis 31(3), 210– 225, http://dx.doi.org/10.1111/j.1467-9892.2010.00657.x.

Liboschik, T. (2016) Modelling count time series following generalized linear models. *PhD Thesis TU Dortmund University, http://dx.doi.org/10.17877/DE290R-17191*.

Liboschik, T., Fokianos, K. and Fried, R. (2017) tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software 82(5), 1– 51, http://dx.doi.org/10.18637/jss.v082.i05*.

Thomas, J.Lonnie, C. (1986) The Economic Design of Control Charts: A Unified Approach. *Journal of Technometrics, 28:1, 3-10.*

Weiss, C. (2010). The INARCH(1) model for overdispersed time series of counts. *Journal of Communications in Statistics Simulation and Computation 39, 1269– 1291*.