The **glarma** Package for Observation Driven Time Series Regression of Counts

William T.M. Dunsmuir, and David J. Scott

Abstract

We review the theory and application of generalised linear autoregressive moving average observation driven models for time series of counts with explanatory variables and describe the estimation of these models using the **glarma** R package. Diagnostic and graphical methods are also illustrated by several examples.

1 Introduction

In the past 15 years there has been substantial progress made in developing regression models with serial dependence for discrete valued response time series such as arise for modeling Bernoulli, binomial, Poisson or negative binomial-counts. In this paper, we consider the GLARMA (generalized linear autoregressive moving average) subclass of observation driven models in detail. Assessing and modelling dependence when the outcomes are discrete random variables is particularly challenging. A major objective of using GLARMA models is the making of inferences concerning regression variables while ensuring that dependence is detected and properly accounted for. GLARMA models are relatively easy to fit and provide an accrssible and rapid way to detect and account for serial dependence in regression modelling of time series.

1.1 Generalized state space models

The GLARMA models considered here are a subclass of generalized state space models for non-Gaussian time series described in Davis *et al.* (1999), Brockwell and Davis (2010) and Durbin and Koopman (2012) for example. A generalized state-space model for a time series $\{Y_t, t = 1, 2, ...\}$ consists of an observation variable and state variable. The model is expressed in terms of conditional probability distributions for the observation and state variables. Such models can be loosely characterized as either parameter driven or observation driven. The observation specification is the same for both models.

For parameter driven models the serial dependence in the state equation is governed by a latent, usually stationary, time series that cannot be observed directly and which evolves independently of past and present values of the observed responses or the covariates. On thte other hand, as the name implies,

in observation driven models, the random component of W_t depends on past observations $\{Y_s, s < t\}$.

Estimation of parameter driven models requires very high dimensional integrals to be evaluated or approximated using asymtotic expansions, simulation methods, numerical integration or all three. Because of this they can be difficult to fit and for routine model building in which many potential regressors need to be considered and evaluated for significance, the parameter driven models for count time series are not yet ready for general use.

On the other hand, the observation driven models considered here are much easier to fit because the likelihood is conditionally specified as a product of conditional distributions which belong to the exponential family and for which the natural parameter is readily calculated via recursion. As a result they are relatively straightforward to apply in practical settings with numerous regressors and long time series.

2 Diagnostics

2.1 Probability integral transformation

To examine the validity of the assumed distribution in the GLARMA model a number of authors have suggested the use of the probability integral transformation (PIT), see for example Czado *et al.* (2009). Although the PIT applies to continuous distributions and the distributions in GLARMA models are discrete, Czado *et al.* (2009) have provided a non-randomized approach which has been implemented in the **glarma** package. There are four functions involved : `glarmaPredProb` calculates conditional predictive probabilities; `glarmaPIT` calculates the non-randomized PIT; `histPIT` plots a histogram of the PIT; and `qqPIT` draws a Q-Q plot of the PIT. If the distribution selected for the model is correct, then the histogram and Q-Q plot should resemble the histogram and Q-Q plot obtained when sampling from the uniform distribution on $[0, 1]$. Of the two plots, the histogram is generally more revealing. Deviations from the expected form of the Q-Q plot are often difficult to discern.

To calculate the conditional predictive probabilities and the PIT the following formulae from Czado *et al.* (2009) are used.

Given the counts $\{y_t\}$, the conditional predictive probability function $F^{(t)}(.|y_t)$ is given by

$$F^{(t)}(u|y_t) = \begin{cases} 0, & u \leq F(y_t - 1), \\ \dfrac{u - F(y_t - 1)}{F(y_t) - F(y_t - 1)}, & F(y_t - 1) \leq u \leq F(y_t), \\ 1, & u > F(y_t). \end{cases} \tag{1}$$

Here $F(y_t)$ and $F(y_t - 1)$ are the upper and lower conditional predictive probabilities respectively.

Then the non-randomized PIT is defined as

$$\bar{F}(u) = \frac{1}{T-1} \sum_{t=2}^{T} F^{(t)}(u|y_t) \tag{2}$$

To draw the PIT histogram, the number of bins, I, is chosen, then the height of the ith bin is

$$f_i = \bar{F}\left(\frac{i}{I}\right) - \bar{F}\left(\frac{i-1}{I}\right). \tag{3}$$

The default number of bins in `histPIT` is 10. To help with assessment of the distribution, a horizontal line is drawn on the histogram at height 1, representing the density function of the uniform distribution on $[0, 1]$.

The Q-Q plot of the PIT plots $\bar{F}(u)$ against u, for $u \in [0, 1]$. The quantile function of the uniform distribution on $[0, 1]$ is also drawn on the plot for reference.

Jung and Tremayne (2011) employ the above diagnostics as well as the randomized version of PIT residuals to compare alternative competing count time series models for several data sets.

2.2 Plots

The plot method for objects of class `"glarma"` produces six plots by default: a time series plot with the observed values of the dependent variable, the fixed effects fit, and the GLARMA fit; an ACF plot of the residuals; a plot of the residuals against time; a normal Q-Q plot; the PIT histogram; and the Q-Q plot for the PIT. Any subset of these six plots can be produce using the `which` argument. For example to omit both of the Q-Q plots (plot 4 and 6), set `which = c(1:3, 5)`. Arguments to the plot method are also provided to change properties of lines in these plots, namely line types, widths, and colours.

3 Examples

There are four example data sets included in the **glarma** package. Sample analyses for all these data sets are provided in either the help pages for the data sets or for the `glarma()` function.

GLARMA models with Poisson counts have appeared previously in the literature, however analyses using the binomial and negative binomial distributions are nobel, so we concentrate on those cases in this section.

3.1 Court Conviction Data

This data set records monthly counts of charges laid and convictions made in Local Courts and Higher Court in armed robbery in New South Wales, Australia, from 1995-2007, see Dunsmuir *et al.* (2008). A description of the columns in the data set is given in Table 1.

The first step is to set up dummy variables for months.

Column	Variable	Description
1	`Data`	Date in monty/year format
2	`Incpt`	Vector of 1s
3	`Trend`	Scaled time trend
4	`Step.2001`	Step change from 2001 onwards
5	`Trend.2001`	Change in trend from 2001 onwards
6	`HC.N`	Monthly number of cases, Higher Court
7	`HC.Y`	Monthly number of convictions, Higher Court
8	`HC.P`	Monthly proportion of convictions, Higher Court
9	`LC.N`	Monthly number of cases, Lower Court
10	`LC.Y`	Monthly number of convictions, Lower Court
11	`LC.P`	Monthly proportion of convictions, Lower Court

Table 1: The court conviction data set.

```
R> data("RobberyConvict")
R> datalen <- dim(RobberyConvict)[1]
R> monthmat <- matrix(0, nrow = datalen, ncol = 12)
R> dimnames(monthmat) <- list(NULL,
+                             c("Jan", "Feb", "Mar", "Apr","May", "Jun",
+                               "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
R> months <- unique(months(strptime(RobberyConvict$Date,
+                                   format = "%m/%d/%Y"),
+                           abbreviate=TRUE))
R> for (j in 1:12) {
+   monthmat[months(strptime(RobberyConvict$Date, "%m/%d/%Y"),
+                   abbreviate = TRUE) == months[j], j] <-1
+ }
R>
R> RobberyConvict <- cbind(rep(1, datalen), RobberyConvict, monthmat)
R> rm(monthmat)
```

Similar analyses can be carried out for both the Lower Court and the Higher Court data. Here we consider only the Lower Court data. The ARIMA component of the model is chosen to be AR(1) and the model for the conviction counts is binomial. A GLM is fitted first to obtain an initial value for the regression coefficients. The initial value of the AR parameter is set at 0. Pearson residuals are used with Newton-Raphson iteration.

First of all the data is prepared for fitting a binomial and the initial GLM fit is obtained.

```
R> y1 <- RobberyConvict$LC.Y
R> n1 <- RobberyConvict$LC.N
R> Y <- cbind(y1, n1-y1)
R> head(Y, 5)
```

```
     y1
[1,]  3 9
[2,]  3 8
[3,]  6 9
[4,]  6 9
[5,]  6 5
```

```
R> glm.LCRobbery <- glm(Y ~ Step.2001 +
+                         I(Feb + Mar + Apr + May + Jun + Jul) +
+                         I(Aug + Sep + Oct + Nov + Dec),
+                       data = RobberyConvict, family = binomial(link = logit),
+                       na.action = na.omit, x = TRUE)
R> summary(glm.LCRobbery, corr = FALSE)


Call:
glm(formula = Y ~ Step.2001 + I(Feb + Mar + Apr + May + Jun +
    Jul) + I(Aug + Sep + Oct + Nov + Dec), family = binomial(link = logit),
    data = RobberyConvict, na.action = na.omit, x = TRUE)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.543  -0.898   0.168   0.801   2.650

Coefficients:
                                    Estimate
(Intercept)                          -0.2568
Step.2001                             0.8232
I(Feb + Mar + Apr + May + Jun + Jul)  -0.3723
I(Aug + Sep + Oct + Nov + Dec)        -0.5007
                                    Std. Error
(Intercept)                             0.1561
Step.2001                               0.0813
I(Feb + Mar + Apr + May + Jun + Jul)    0.1619
I(Aug + Sep + Oct + Nov + Dec)          0.1655
                                    z value Pr(>|z|)
(Intercept)                           -1.65   0.0998
Step.2001                             10.12   <2e-16
I(Feb + Mar + Apr + May + Jun + Jul)  -2.30   0.0215
I(Aug + Sep + Oct + Nov + Dec)        -3.03   0.0025

(Intercept)                         .
Step.2001                           ***
I(Feb + Mar + Apr + May + Jun + Jul) *
I(Aug + Sep + Oct + Nov + Dec)       **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 327.48  on 149  degrees of freedom
Residual deviance: 212.12  on 146  degrees of freedom
AIC: 684.8

Number of Fisher Scoring iterations: 4
```

We can print the regression coeffcients for this model in a more elegant form in Table 2.

```
glmCoeff <- summary(glm.LCRobbery)$coefficients
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.26	0.16	-1.65	0.10
Step.2001	0.82	0.08	10.12	0.00
I(Feb + Mar + Apr + May + Jun + Jul)	-0.37	0.16	-2.30	0.02
I(Aug + Sep + Oct + Nov + Dec)	-0.50	0.17	-3.03	0.00

Table 2: Regression coeffcients

Then the GLARMA model is fitted.

```
R> X <- glm.LCRobbery$x
R> colnames(X)[3:4] <- c("Feb-Jul","Aug-Dec")
R> head(X, 5)

  (Intercept) Step.2001 Feb-Jul Aug-Dec
1           1         0       0       0
2           1         0       1       0
3           1         0       1       0
4           1         0       1       0
5           1         0       1       0

R> glarmamod <- glarma(Y, X, phiLags = c(1), type = "Bin", method = "NR",
+                      maxit = 100, grad = 1e-6)
R> summary(glarmamod)


Call: glarma(y = Y, X = X, type = "Bin", method = "NR", phiLags = c(1),
    maxit = 100, grad = 1e-06)

Pearson Residuals:
```

```
   Min      1Q  Median      3Q     Max
-2.446  -0.816   0.134   0.730   2.480

Autoregressive Coefficients:
      Estimate Std.Error z-ratio Pr(>|z|)
phi_1   0.0818    0.0330    2.48    0.013 *

Linear Model Coefficients:
            Estimate Std.Error z-ratio Pr(>|z|)
(Intercept)  -0.2747    0.1571   -1.75   0.0804 .
Step.2001     0.8220    0.0957    8.59   <2e-16 ***
Feb-Jul      -0.3568    0.1598   -2.23   0.0256 *
Aug-Dec      -0.5004    0.1633   -3.06   0.0022 **

    Null deviance: 327.48  on 149  degrees of freedom
Residual deviance: 198.91  on 145  degrees of freedom
AIC: 680.7

Number of Newton Raphson iterations: 4

LRT and Wald Test:
Alternative hypothesis: model is a GLARMA process
Null hypothesis: model is a GLM with the same regression structure
          Statistic p-value
LR Test        6.11   0.013 *
Wald Test      6.14   0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe that the regression coeffcients for the GLARMA model are quite similar to those for the GLM model. In particular, the step change in 2001 is highly significant. The likelihood ratio and Wald tests both suggest the need to deal with autocorrelation.

```
par(mar = c(4, 4, 3, 0.1), cex.lab = 0.95, cex.axis = 0.9, mgp = c(2,
    0.7, 0), tcl = -0.3, las = 1)
plot(glarmamod)
```

In the diagnostic plots shown in Figure 1, the ACF plot shows little residual autocorrelation, and the Q-Q plot of the residuals shows reasonable conformity with normality. However the PIT histogram suggests that the binomial model for the counts is not appropriate for this data.

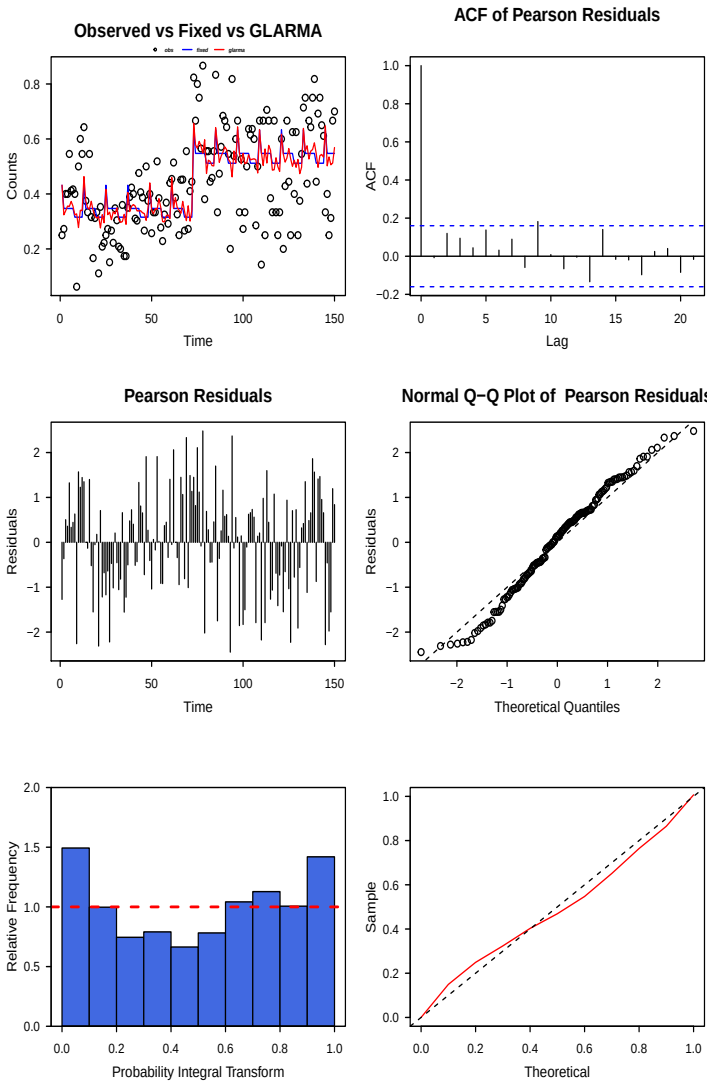

Figure 1: Diagnostic plots for the court conviction model.

References

Brockwell PJ, Davis RA (2010). *Introduction to Time Series and Forecasting*. 2nd edition edition. Springer-Verlag, New York, NY.

Czado C, Gneiting T, Held L (2009). "Predictive Model Assessment for Count Data." *Biometrics*, **65**(4), 1254–1261. doi:10.1111/j.1541-0420.2009.01191.x. URL `http://dx.doi.org/10.1111/j.1541-0420.2009.01191.x`.

Davis RA, Dunsmuir WT, Wang Y (1999). "Modeling Time Series of Count Data." In S Ghosh (ed.), *Asymptotics, Nonparametrics, and Time Series*, volume 158 of *Statistics Textbooks and Monographs*, pp. 63-114. Marcel Dekker, New York, NY.

Dunsmuir WT, Tran CD, Weatherburn D, Wales NS (2008). *Assessing the Impact of Mandatory DNA Testing of Prison Inmates in NSW on Clearance, Charge and Conviction Rates for Selected Crime Categories*. NSW Bureau of Crime Statistics and Research.

Durbin J, Koopman SJ (2012). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Jung RC, Tremayne AR (2011). "Useful Models for Time Series of Counts or Simply Wrong Ones?" *Advances in Statistical Analysis*, **95**(1), 59–91.