The glarma Package for Observation Driven Time Series Regression of Counts

William T.M. Dunsmuir
University of New South Wales

David J. Scott
University of Auckland

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.

Keywords: observation driven count time series, generalized linear arma models, **glarma**, R.

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 modelling 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 accessible 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, Dunsmuir, and Wang (1999), Brockwell and Davis (2010) and Durbin and Koopman (2012) for example. A generalized state-space model for a time series $\{Y_t, t = 1, 2, \ldots\}$ 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 the 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 asymptotic 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.

The outline of the remainder of the paper is as follows. Section 2 provides the necessary theoretical background for GLARMA models. It describes the various combinations of model elements (response distributions, dependence structures and predictive residuals) that are currently supported in the **glarma** package. Options for initializing and obtaining maximum likelihood estimates using a form of Fisher scoring or Newton-Raphson iterations are described. Issues of parameter identifiability and convergence properties of GLARMA models and the maximum likelihood estimates are also reviewed to guide users in the application of these models. Section 3 describes the various modelling functions available in the package. Section 4 describes the built-in model diagnostic procedures and plotting functions. Section 5 provides several examples illustrating the use of the package on real data sets.

2. Theory of GLARMA models

The **glarma** package provides functionality for estimating regression relationships between a vector of regressors (covariates, predictors) and a discrete valued response. In time series regression modelling it is typically the case that there is serial dependence. This package models the serial dependence using the GLARMA class of observation driven models and provides valid inference for the regression model components.

Let there be N consecutive times at which the response and regressor series are observed. The response series are observations on the random variables $\{Y_t : t = 1, \ldots, N\}$ and associated with these are K-dimensional vectors of regressors x_t observed also for $t = 1, \ldots, N$. We let $\mathcal{F}_t = \{Y_s : s < t, x_s : s \leq t\}$ be the past information on the response series and the past and present information on the regressors. In general the conditional distribution of Y_t given $\mathcal{F}_t$ is given in exponential family form as

$$f(y_t|W_t) = \exp\{y_t W_t - a_t b(W_t) + c_t\} \tag{1}$$

where a_t and c_t are sequences of constants possibly depending on the observations y_t. Information in $\mathcal{F}_t$ is summarized in the state variable W_t. Details are provided below for specific distributions available in **glarma**.

Note that (1) is not the fully general form of the exponential family (see McCullagh and Nelder 1989) in that it does not include an over-dispersion parameter and the canonical link is used. It follows from (1) that the conditional means and variances of the responses are $\mu_t := E(Y_t|W_t) = a_t \dot{b}(W_t)$ and $\sigma_t^2 := \text{var}(Y_t|W_t) = a_t \ddot{b}(W_t)$. The negative binomial case is special—see below.

Observation driven models take various forms (see Benjamin, Rigby, and Stasinopoulos 2003, for a general discussion). Here we focus on the case where the state vector in (1) is of the general form

$$W_t = x_t^T \beta + Z_t. \tag{2}$$

In addition to the regression parameters β we assume that there are other parameters ψ which specify the process $\{Z_t\}$ as discussed below.

2.1. GLARMA dependence structure

Serial dependence in the response process can be introduced via Z_t in the state process using linear combinations of past predictive residuals e_t as

$$Z_t = \sum_{j=1}^{\infty} \gamma_j e_{t-j} \tag{3}$$

The predictive residuals are defined as

$$e_t = \frac{Y_t - \mu_t}{\nu_t}. \tag{4}$$

for some scaling sequence $\{\nu_t\}$—see Section 2.3 for choices currently supported. Note that these are martingale differences, hence are zero mean and uncorrelated. When ν_t is set to the conditional standard deviation of Y_t the e_t are also unit variance, hence are weakly stationary white noise.

One parsimonious way in which to parameterize the infinite moving average weights γ_j in (3), is to allow them to be the coefficients in an autoregressive-moving average filter. Specifically, set

$$\sum_{j=1}^{\infty} \gamma_j \zeta^j = \theta(\zeta)/\phi(\zeta) - 1,$$

where $\phi(\zeta) = 1 - \phi_1\zeta - \cdots - \phi_p\zeta^p$ and $\theta(\zeta) = 1 + \theta_1\zeta + \cdots + \theta_q\zeta^q$ are the respective autoregressive and moving average polynomials of the ARMA filter, each having all zeros outside the unit circle. It follows that $\{Z_t\}$ satisfies the ARMA-like recursions,

$$Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q} \theta_i e_{t-i}. \tag{5}$$

The $\{Z_t\}$ defined in this way can be thought of as the best linear predictor of a stationary invertible ARMA process with driving noise specified by the sequence $\{e_t\}$ of scaled deviations of count responses from their conditional mean given the past responses and the past and current regressors. This specification allows recursive calculation (in time t) of the state equation. The model is referred to as a GLARMA model (see (Davis, Dunsmuir, and Streett 2003)). Shephard (1995) provides the first example of such a model class.

2.2. Response distributions

Specific examples of exponential family members of the form (1) that are currently supported are Poisson, negative binomial and binomial which includes Bernoulli as a special case.

Poisson: Here $a_t \equiv 1$, $b(W_t) = \exp(W_t)$, $c_t = -\log y_t!$ and the canonical link is $g(\mu) = \ln(\mu)$. Note that $\mu_t = \exp(W_t)$ and $\sigma_t^2 = \exp(W_t)$.

Binomial/Bernoulli: Let the number of trials at time t be m_t and $\pi_t = P(Y_t = 1|W_t)$. Then $a_t = m_t$, $b(\theta) = \ln(1 + \exp(W_t))$ and $c_t = \log \binom{m_t}{y_t}$. The canonical link is the logit so that $W_t = \log(\pi_t/(1 - \pi_t))$. Note that $\mu_t = m_t\pi_t$ and $\sigma_t^2 = m_t\pi_t(1 - \pi_t)$. The Bernoulli case has $m_t \equiv 1$.

Negative binomial: Let $\mu_t = \exp(W_t)$. The **glarma** package uses the negative binomial density in the form

$$f(y_t|W_t, \alpha) = \frac{\Gamma(\alpha + y_t)}{\Gamma(\alpha)\Gamma(y_t + 1)} \left[\frac{\alpha}{\alpha + \mu_t}\right]^{\alpha} \left[\frac{\mu_t}{\alpha + \mu_t}\right]^{y_t}. \tag{6}$$

Note that $\mu_t = \exp(W_t)$ and $\sigma_t^2 = \mu_t + \mu_t^2/\alpha$. As $\alpha \to \infty$ the negative binomial density converges to the Poisson density. Also note that if α is known, this density can be put in the one parameter exponential family with appropriate definitions of θ_t, $b(\theta_t)$, $a(\psi)$, $c_t(y_t, \psi)$. If α is not known then (6) is not a member of the one parameter exponential family.

2.3. Types of GLARMA residuals

GLARMA predictive residuals are of the form (4) where $\nu(W_t)$ is a scaling function. Currently several choices for this are supported.

Pearson Scaling Here $\nu_t = \nu_{P,t}$ where

$$\nu_{P,t} = [a_t\ddot{b}(W_t)]^{0.5}$$

in which case Pearson residuals result.

Score-type Scaling These replace conditional standard deviation by conditional variances

$$\nu_{S,t} = a_t\ddot{b}(W_t)$$

resulting in the 'score-type' residuals used in Creal, Koopman, and Lucas (2008).

Identity Scaling A third option, which allows some form of the BARMA (binary ARMA) models considered in Wang and Li (2011) to be fit is to use no scaling with

$$\nu_{I,i} = 1.$$

For the Poisson response distribution GLARMA model, failure to scale by the variance or standard deviation function will lead to unstable Poisson means (that diverge to infinity or collapse to zero as an absorbing state for instance) and existence of stationary and ergodic solutions to the recursive state equation is not assured—see Davis *et al.* (1999), Davis *et al.* (2003) and Davis, Dunsmuir, and Streett (2005) for details. For the binomial situation this lack of scaling should not necessarily lead to instability in the success probability as time evolves since the success probabilities, p_t, and observed responses, Y_t, are both bounded between 0 and 1. Thus degeneracy can only arise if the regressors x_t become unbounded from below or above. As recommended in Davis *et al.* (1999) temporal trend regressors should be scaled using a factor relating to sample size n.

2.4. The GLARMA likelihood

Given n successive observations $\{y_t : t = 1, \ldots, n\}$ on the response series the likelihood is constructed as the product of conditional densities of Y_t given $\mathcal{F}_t$. The state vector W_t at each time embodies these conditioning variables and so the log likelihood is given by

$$l(\delta) = \sum_{t=1}^{n} \log f_{Y_t|W_t}(y_t|W_t; \delta). \tag{7}$$

For the Poisson and binomial response distributions the log-likelihood (7) is

$$l(\delta) = \sum_{t=1}^{n} \{y_t W_t(\delta) - a_t b(W_t(\delta)) + c_t\} \tag{8}$$

where $\delta = (\beta, \phi, \theta)$.

For the negative binomial response distribution the log-likelihood is more complicated because the shape parameter α also has to be estimated along with β, ϕ and θ. We then let $\delta = (\beta, \phi, \theta, \alpha)$.

Note that e_t in (4), the Z_t in (5) and thus the W_t in (2) are functions of the unknown parameter δ and hence need to be recomputed for each iteration of the likelihood optimization. Thus in order to calculate the likelihood and its derivatives, recursive expressions are needed to calculate e_t, Z_t and W_t as well as their first and second partial derivative with respect to δ. Expressions for these recursive formulae are available in Davis *et al.* (2005) for the Poisson case. Corresponding formulae for the binomial case were derived in Lu (2002) and for the negative binomial case in Wang (2004). The essential computational cost is in the recursions for Z_t and W_t and their first and second derivative with respect to δ. Fortunately, these require identical code for the various response distributions and definitions of predictive residuals e_t.

For calculation of the Z_t in (5), initializing conditions for the recursions must be used. The current implementation in **glarma** is to set $e_t = 0$ and $Z_t = 0$ for $t \leq 0$ ensuring that the conditional and unconditional expected values of e_t are zero for all t.

The likelihood is maximized from a suitable starting value of the parameter δ using a version of Fisher scoring iteration or by Newton-Raphson iteration. For a given value of δ let the vector of first derivatives with respect to δ of the log-likelihood (7) be

$$d(\delta) = \frac{\partial l(\delta)}{\partial \delta}$$

and the second derivative matrix be

$$D_{NR}(\delta) = \frac{\partial^2 l(\delta)}{\partial \delta \partial \delta^\top}, \tag{9}$$

where the matrix of second derivatives of the log-likelihood is (in the Poisson and binomial response cases) given by

$$D_{NR}(\delta) = \sum_{t=1}^{n} [y_t - a_t \dot{b}(W_t)] \frac{\partial^2 W_t}{\partial \delta \partial \delta^\top} - \sum_{t=1}^{n} a_t \ddot{b}(W_t) \frac{\partial W_t}{\partial \delta} \frac{\partial W_t}{\partial \delta^\top}. \tag{10}$$

and $\dot{b}(u)$ and $\ddot{b}(u)$ are the first and second derivatives respectively of the function $b(u)$ with respect to the argument u.

Using the fact that, at the true parameter value δ, $E[y_t - a_t\dot{b}(W_t)|\mathcal{F}_t] = 0$ the expected value the first summation in (10) is zero and hence the expected value of the matrix of second derivatives is $E[D_{FS}(\delta)]$ where

$$D_{FS}(\delta) = -\sum_{t=1}^{n} a_t\ddot{b}(W_t)\frac{\partial W_t}{\partial \delta}\frac{\partial W_t}{\partial \delta^\top}. \tag{11}$$

Note also that due to the martingale difference property of the predictive residuals we also have $E[D_{NR}(\delta)] = -E[d(\delta)d(\delta)^\top]$. While these expectations cannot be computed in closed form, expression (11) requires first derivatives only and is used in package **glarma** as the basis for the approximate Fisher scoring method.

Thus, if $\delta^{(k)}$ is the parameter vector at the current iterate k, the Newton-Raphson updates proceed using

$$\delta^{(k+1)} = \delta^{(k)} - D_{NR}(\delta^{(k)})^{-1}d(\delta^{(k)}) \tag{12}$$

and the approximate Fisher scoring updates use D_{FS} in place of D_{NR}

Given a specified tolerance TOL, iterations continue until the largest gradient of the log-likelihood satisfies $\max_i |d_i(\delta^{(k)}|) \leq TOL$ or a maximum number of iterations MAXITER is surpassed. At termination we let $\hat{\delta} = \delta^{(k+1)}$ and call this the "maximum likelihood estimate" of δ.

By default, the iterations in (12) are initialized using the generalized linear model (GLM) estimates for β and zero initial values for the autoregressive moving average terms. For the negative binomial case β and α are initialized using a call to `glm.nb()` from the package **MASS** (see Venables and Ripley (2002)). Convergence in the majority of cases is rapid. Users may optionally specify initial parameter values of their own choice.

2.5. Parameter identifiability

The GLARMA component Z_t of the state variable given in (5) can be rewritten as

$$Z_t = \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{i=1}^{\tilde{q}} \tilde{\theta}_i e_{t-i}. \tag{13}$$

where $\tilde{q} = \max(p, q)$ and

1. If $p \leq q$, $\tilde{\theta}_j = \theta_j + \phi_j$ for $j = 1, \ldots, p$ and $\tilde{\theta}_j = \theta_j$ for $j = p+1, \ldots, q$.
2. If $p > q$, $\tilde{\theta}_j = \theta_j + \phi_j$ for $j = 1, \ldots, q$ and $\tilde{\theta}_j = \phi_j$ for $\jmath = q+1, \ldots, p$.

When pre-observation period values are set to zero (that is $Z_t = 0$ for $t \leq 0$ and $e_t = 0$ for $t \leq 0$) then if and only if $\tilde{\theta}_j = 0$ for $j = 1, \ldots, \tilde{p}$ the recursion (13) would result in $Z_t = 0$ for all t and hence there is no serial dependence in the GLARMA model. This is is equivalent to $\phi_j = -\theta_j$ for $j = 1, \ldots, p$ and $\theta_j = 0$ for $j = p+1, \ldots, \tilde{q}$.

Consequently a null hypothesis of no serial dependence requires only these constraints on the θ and ϕ parameters. In some situations this means that under the null hypothesis of no serial

dependence there are nuisance parameters which cannot be estimated. This has implications for convergence of the iterations required to optimize the likelihood and on testing that there is no serial dependence in the observations (other than induced by the regression component $x_t^\top \beta$).

When $p > 0$ and $q = 0$ (equivalent to an ARMA(p, p) specification with constraint $\theta_j = \phi_j$ or a pure MA with $p = 0$ and $q > 0$ then identification issues do not arise and the hypothesis of no serial dependence corresponds to the hypothesis that $\phi_j = 0$ for $j = 1, \ldots, p$ in the first case and $\theta_j = 0$ for $j = 1, \ldots, q$ in the second case. The provided likelihood ratio and Wald tests (see Section 4.1 for further details) will have an asymptotic chi-squared distribution with correct degrees of freedom.

In cases where $p > 0$ and $q > 0$ some caution is advised when fitting models and testing that serial dependence is not present. To simplify the discussion we focus on the case where $p = q$;

1. If there is no serial dependence in the observations but $p = q > 0$ is specified then there is a strong possibility that the likelihood optimization for this overspecified model will not converge because the likelihood surface will be 'ridge-like' along the line where $\phi_j = -\theta_j$. This issue is classical for standard ARMA models. Similarly if the degree of serial dependence is of lower order than that specified for the GLARMA model identifiability issues and lack of convergence of the likelihood optimizing recursions is likely to occur. Following from this it is highly recommended that users start with low orders for p and q and initially avoid specifying them to be equal. Once stability of estimation is reached for a lower order specification increasing the values of p or q could be attempted. Lack of identifiability typically manifests itself in the matrix of second derivatives D_{NR} or the approximate Fisher scoring version D_{FS} becoming close to singular or even non-positive definite. The state variable W_t can also degenerate to $\pm\infty$ for which an error code in the output from the `glarma()` call is provided.

2. The likelihood ratio test that there is no serial dependence versus the alternative that there is GLARMA like serial dependence with $p = q > 0$ will not have a standard chi-squared distribution because the parameters ϕ_j for $j = 1, \ldots, p$ are nuisance parameters which cannot be estimated under the null hypothesis. Testing methods such as proposed in Hansen (1996) need to be developed for this situation.

2.6. Stochastic properties of GLARMA models

Means, variances and autocovariances for the state process $\{W_t\}$ can be readily derived using the definition of Z_t in (3)—see (Davis *et al.* 1999). For the Poisson response case the corresponding means, variance and autocovariances for the count response series $\{Y_t\}$ can be derived approximately. Additionally an approximate interpretation of the regression coefficients β can be given—see (Davis *et al.* 2003). Similar results could be derived for the negative binomial response case. For binomial and Bernoulli responses, calculation of means, variances, autocovariances for the response series and interpretation of regression coefficients is not straightforward. This is a typical issue for interpretation of random effects models and transition models in the binomial or Bernoulli case—see Diggle, Heagerty, Liang, and Zeger (2002) for example.

To date the stationarity and ergodicity properties of the GLARMA model are only partially understood. These properties are important in order to ensure that the process is capable of

generating sample paths that do not degenerate to zero or do not explode as time progresses, as well as for establishing the large sample distributional properties of estimates of the parameters. Davis *et al.* (2003) provide partial results for perhaps the simplest of all possible models for Poisson responses specified with $p = 0$, $q = 1$ and $x_t^\top \beta = \beta$. Results for simple examples of the stationary Bernoulli case are given in Streett (2000).

2.7. Fitted values

There are two concepts of fitted values currently supported for the GLARMA model. The first is defined as the estimated conditional mean function $\hat{\mu}_t$ at time t calculated using the maximum likelihood estimates $\hat{\delta}$. Thus

$$\hat{\mu}_t = m_t \dot{b}(x_t^\top \hat{\beta} + \hat{Z}_t) \tag{14}$$

where $\hat{Z}_t$ are calculated using $\hat{\delta}$ in (5). These fitted values combine the regression fit (fixed effects) together with the contribution from weighted sums of past estimated predictive residuals.

Because for GLARMA models the unconditional mean function is difficult to obtain exactly in all cases an estimated unconditional mean function of t is not provided. Instead, for the second concept of fitted values, the fitted value from the regression term only is suggested as a guide to the fit without the effect of random variation due to Z_t. This is defined to be

$$\tilde{\mu}_t = m_t \dot{b}(x_t^\top \hat{\beta}) \tag{15}$$

We refer to this as the "fixed effects fit" in plotting functions below. Note that this is not an estimate of the unconditional mean even in the Poisson case (arguably the most tractable for this calculation)—the theoretical unconditional mean for this case is approximated by $\exp(x_t^\top \beta + \nu^2/2)$ where $\nu^2 = \sum_{l=1}^{\infty} \gamma_l^2$—see Davis *et al.* (2003) for details. A similar calculation for the binomial case is not available. Hence, in view of these theoretical constraints, the use of the fixed effects fit seems a simple and sensible alternative to the conditional mean $\hat{\mu}_t$ given by (14).

2.8. Distribution theory for likelihood estimation

For inference in the GLARMA model it is assumed that the central limit theorem holds so that

$$\hat{\delta} \overset{d}{\approx} N(\delta, \hat{\Omega}) \tag{16}$$

where the approximate covariance matrix is estimated by

$$\hat{\Omega} = -D_{NR}(\hat{\delta})^{-1} \tag{17}$$

in the case of Newton-Raphson and similarly with D_{NR} replaced by D_{FS} in the case of Fisher scoring. Thus a standard error for the maximum likelihood estimates of the ith component of δ is computed using $\hat{\Omega}_{ii}^{1/2}$.

There have been a number of claims in the literature concerning a central limit theorem for models of this type. However all of these make assumptions concerning convergence of key quantities all of which require the ergodicity to be established which has not been done in generality as yet. The central limit theorem for the maximum likelihood parameter estimates

is rigorously established only in the stationary Poisson response case in Davis *et al.* (2003) and in the Bernoulli stationary case in Streett (2000). Simulation results are also reported in Davis *et al.* (1999, 2003) for non-stationary Poisson models. Other simulations not reported in the literature support the supposition that the estimates $\hat{\delta}$ have a multivariate normal distribution for large samples for a range of regression designs and for the various response distributions considered here.

A central limit theorem for the maximum likelihood estimators is currently not available for the general model. Regardless of the technical issues involved in establishing a general central limit theorem the above approximate result seems plausible since, for these models the log-likelihood is a sum of elements in a triangular array of martingale differences.

For nested models likelihood ratio test statistics can be calculated and compared to the assumed chi-squared asymptotic distribution in the usual way. The above asymptotic result can be used to obtain an approximate chi-squared distribution for a Wald test that subsets of δ take specified values. Let $\delta^{(1)}$ specify a subset of δ that is hypothesized to take a specific value $\delta_0^{(1)}$. The Wald test is constructed as

$$W^2 = [\hat{\delta}^{(1)} - \delta_0^{(1)}]^\top [\hat{\Omega}^{(1)}]^{-1} [\hat{\delta}^{(1)} - \delta_0^{(1)}] \tag{18}$$

where $\hat{\Omega}^{(1)}$ is the submatrix corresponding to $\delta_0^{(1)}$ of the estimated asymptotic covariance matrix of (17).

Further details on implementation of these tests in **glarma** are given in Section 4.1.

3. Modelling functions

There are seven modelling functions for fitting GLARMA models, falling into three groups:

Poisson: `glarmaPoissonPearson()` and `glarmaPoissonScore()`.

Binomial: `glarmaBinomialIdentity()`, `glarmaBinomialPearson()` and `glarmaBinomialScore()`.

Negative Binomial: `glarmaNegBinPearson()` and `glarmaNegBinScore()`.

The second component of the name indicates the distribution used for the counts and the third component the residuals used in the fitting routine. A call to `glarma()` results in a call to the appropriate fitting routine, as determined by the values of the arguments `type`, and `residuals` supplied to the `glarma()` call. Pearson residuals are used by default. Two iterative methods are available for the optimization of the log-likelihood, Fisher Scoring (`method = "FS"`) and Newton-Raphson (`method = "NR"`), the default method being Fisher Scoring. The object returned by any of the fitting routines is of class `"glarma"`.

To specify the model in a call to `glarma()`, the response variable is given by the argument `y`, and the matrix of predictors for the regression part of the model is given by the argument `X`. The matrix `X` must include a column of ones to enable the fitting of a mean term in the regression component of the model. Initial values can be given for the coefficients in the regression component using the argument `beta`. If no initial values are provided, a call is made to the corresponding generalized linear model to obtain initial regression coefficient values.

The ARMA component of the model is specified using the arguments `phiLags` and `phiInit` (for the AR terms) and `thetaLags` and `thetaInit` (for the MA terms). For both the AR and MA terms, the first argument of the pair of arguments specifies the orders of the lags which are to be included in the model, and the second argument the initial values of the coefficients for those lags.

When the counts are modeled using the negative binomial distribution, there is an additional parameter, the shape parameter of the negative binomial, designated as α in the GLARMA model. This parameter is called θ in the function `glm.nb()` from the package **MASS**, but for GLARMA models θ refers to the moving average terms in the ARMA component of the model. An initial value for α can be provided using the argument `alphaInit`. If no initial value is provided, a call is made to `glm.nb()` from **MASS**. An initial value for the call to `glm.nb()` can be supplied by giving a value to the argument `alpha` of `glarma()`. The default value for `alpha` is 1.

Because the GLARMA model is fitted using numerical non-linear optimization, non-convergence is a possibility. Two error codes are included in the object returned by the `glarma()` to alert users to numerical problems with fitting. If the Fisher Scoring or Newton-Raphson iterations fail to converge, `errCode` will be set to 1. This can result from non-identifiability of the ARMA component of the model such as when the degrees and lags of both the AR and MA components are specified to be the same, as discussed in Section 2.5. It is possible that for certain values of the ARMA parameters the recursions calculating $\{W_t\}$ diverge to $\pm\infty$. In that case the value of `WError` will be set to 1 allowing the user to check for this condition when the likelihood optimization fails to converge.

Once a fitted model object has been obtained, there are accessor functions available using S3 methods to extract the coefficients (`coef()`, or the alias `coefficients()`), the fitted values (`fitted()` or the alias `fitted.values()`), the residuals (`residuals()` or the alias `resid()`), the model frame (`model.frame()`), the number of observations (`nobs()`), the log-likelihood (`logLik()`), and the AIC (`extractAIC()`). These are standard implementations of these methods with the exception of `coef()`. This method takes an argument `types` which allows the extraction of the ARMA coefficients (`types = "ARMA"`), or the regression coefficients (`types = "beta"`), or both sets of coefficients (`types = "all"`), the default.

Other S3 methods available for an object of class `"glarma"` are `print`, `summary`, `print.summary`, and `plot`.

4. Diagnostics

4.1. Likelihood ratio and Wald tests

In **glarma**, the likelihood ratio test and the Wald test tests that the serial dependence parameters $\psi = (\phi^\top, \theta^\top)^\top$ are all equal to zero (that is tests $H_0 : \psi = 0$ versus $H_a : \psi \neq 0$) are provided by the function `likTests()`, which operates on an object of type "glarma". The likelihood ratio test compares the likelihood of the fitted GLARMA model with the likelihood of the GLM model with the same regression structure. The same null hypothesis applies to the Wald test, which is based on the Wald statistic defined in (18). Values of both statistics are compared to the chi-squared distribution with degrees of freedom given by the number of ARMA parameters. These degrees of freedom and associated chi-squared p values are correct

under the situations discussed in Section 2.5.

Package users may also construct their own tailor made likelihood ratio tests by using the reported log-likelihood (`logLik()`) for the two models under comparison and Wald tests W^2 in (18) using the appropriate submatrix of the reported estimated covariance matrix in (17) available as `glarmamod$cov`.

4.2. 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, Gneiting, and Held (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

$$F^{(t)}(u|y_t) = \begin{cases} 0, & u \leq F(y_t - 1), \\ \dfrac{u - F(y_t - 1)}{F(y_t) - F(y_t - 1)}, & F(y_t - 1) \leq u \leq F(y_t), \\ 1, & u > F(y_t). \end{cases} \tag{19}$$

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

$$\bar{F}(u) = \frac{1}{T-1} \sum_{t=2}^{T} F^{(t)}(u|y_t) \tag{20}$$

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

$$f_i = \bar{F}\left(\frac{i}{I}\right) - \bar{F}\left(\frac{i-1}{I}\right). \tag{21}$$

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.

4.3. 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 (plots 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.

5. 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 novel, so we concentrate on those cases in this section.

5.1. Asthma data

This data set arose from a single hospital (at Campbelltown, as part of a larger study into the relationship between atmospheric pollution and the number of asthma cases presenting at emergency departments in the South West region of Sydney, Australia, see Davis *et al.* (2003). A description of the columns in the data set is given in Table 5.1

Column	Variable	Description
1	`Count`	Daily asthma counts
2	`Intercept`	Vector of 1s
3	`Sunday`	Dummy variable for Sundays
4	`Monday`	Dummy variable for Mondays
5	`CosAnnual`	$\cos(2\pi t/365)$, annual cosine term
6	`SinAnnual`	$\sin(2\pi t/365)$, annual sine term
7	`H7`	Scaled, lagged and smoothed humidity
8	`NO2max`	Maximum daily nitrogen oxide
9–16	`T1.1990-T2.1993`	Smooth shapes to capture school terms in each year

Table 1: The asthma data set

We fit a model with a moving average term at lag 7 with negative binomial counts. The initial values of the regression coefficients are found by fitting the corresponding GLM model, and the initial value of the shape parameter, α of the negative binomial distribution is taken as 0. Pearson residuals are used and fitting is by Newton-Raphson.

```
> data(Asthma)
> y <- Asthma[, 1]
> X <- as.matrix(Asthma[, 2:16])
> glarmamod <- glarma(y, X, thetaLags = 7, type = "NegBin", method = "NR",
+                     residuals = "Pearson", alphaInit = 0,
+                     maxit = 100, grad = 1e-6)
> glarmamod


Call: glarma(y = y, X = X, type = "NegBin", method = "NR", residuals = "Pearson",
    thetaLags = 7, alphaInit = 0, maxit = 100, grad = 1e-06)

Negative Binomial Parameter:
alpha
37.19

Autoregressive Coefficients:
theta_7
0.04392

Linear Model Coefficients:
Intercept     Sunday     Monday  CosAnnual  SinAnnual
  0.58397    0.19455    0.22999   -0.21450    0.17728
       H7     NO2max    T1.1990    T2.1990    T1.1991
  0.16843   -0.10404    0.19903    0.13087    0.08587
  T2.1991    T1.1992    T2.1992    T1.1993    T2.1993
  0.17082    0.25276    0.30572    0.43607    0.11412

Degrees of Freedom: 1460 Total (i.e. Null);  1444 Residual
Null Deviance: 1990
Residual Deviance: 1443
AIC: 4874

> summary(glarmamod)


Call: glarma(y = y, X = X, type = "NegBin", method = "NR", residuals = "Pearson",
    thetaLags = 7, alphaInit = 0, maxit = 100, grad = 1e-06)

Pearson Residuals:
   Min      1Q  Median      3Q     Max
-1.849  -0.741  -0.175   0.609   6.178

Negative Binomial Parameter:
      Estimate Std.Error z-ratio Pr(>|z|)
alpha     37.2      25.4    1.46     0.14
```

```
Autoregressive Coefficients:
        Estimate Std.Error z-ratio Pr(>|z|)
theta_7   0.0439    0.0194    2.27    0.023 *

Linear Model Coefficients:
          Estimate Std.Error z-ratio Pr(>|z|)
Intercept   0.5840    0.0633    9.22  < 2e-16 ***
Sunday      0.1946    0.0576    3.38  0.00073 ***
Monday      0.2300    0.0564    4.08  4.6e-05 ***
CosAnnual  -0.2145    0.0397   -5.41  6.3e-08 ***
SinAnnual   0.1773    0.0415    4.27  2.0e-05 ***
H7          0.1684    0.0563    2.99  0.00279 **
NO2max     -0.1040    0.0339   -3.07  0.00216 **
T1.1990     0.1990    0.0585    3.41  0.00066 ***
T2.1990     0.1309    0.0590    2.22  0.02648 *
T1.1991     0.0859    0.0675    1.27  0.20306
T2.1991     0.1708    0.0595    2.87  0.00410 **
T1.1992     0.2528    0.0567    4.46  8.2e-06 ***
T2.1992     0.3057    0.0510    5.99  2.1e-09 ***
T1.1993     0.4361    0.0523    8.33  < 2e-16 ***
T2.1993     0.1141    0.0627    1.82  0.06868 .

    Null deviance: 1989.9  on 1460  degrees of freedom
Residual deviance: 1442.6  on 1444  degrees of freedom
AIC: 4874

Number of Newton Raphson iterations: 6

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        7.05  0.0079 **
Wald Test      5.15  0.0233 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We note that virtually all the regression terms in the model are significant, most being highly significant. The moving average term is significant and both the tests indicate that there is a need to fit a GLARMA model rather than a simple GLM. The value of α is quite large, suggesting that a Poisson model might provide adequate fit.

The plot method for an object of class `"glarma"` shows six plots by default: a time series plot with observed values of the dependent variable, fixed effects fit, and GLARMA fit; an ACF plot of residuals; a plot of residuals against time; a normal Q-Q plot; the PIT histogram; and the uniform Q-Q plot for the PIT. As an example, in Figure 1, we show just four of these plots. Since the default title for the PIT histogram is too long for the available space we use

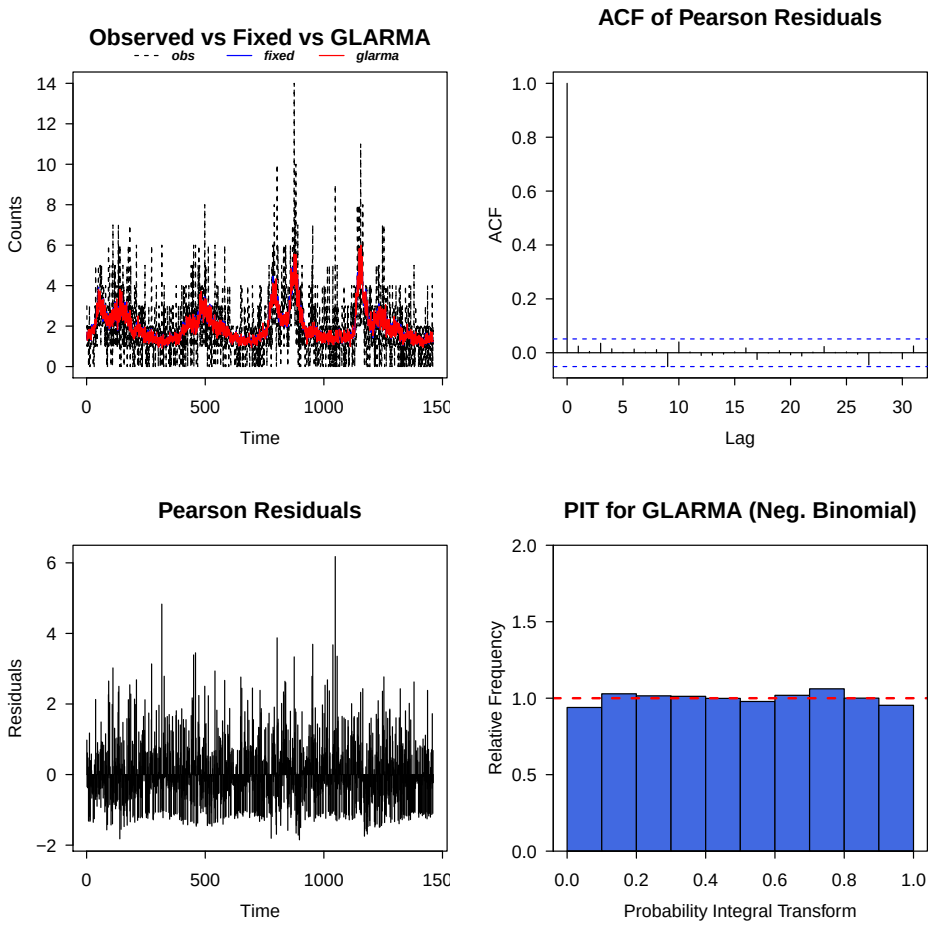

Figure 1: Diagnostic plots for the asthma model

the `titles` argument to abbreviate it.

```
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, which = c(1, 2, 3, 5), titles = list(NULL, NULL,
    NULL, "PIT for GLARMA (Neg. Binomial)"))
```

The ACF plot indicates that the model has dealt adequately with any serial correlation present, and the PIT histogram suggests that the negative binomial provides a suitable model for the counts.

5.2. 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, Tran, Weatherburn, and Wales (2008). A description of the columns in the data set is given in Table 5.2.

Column	Variable	Description
1	`Date`	Date in month/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 2: The court conviction data set

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

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

```
> ### Prepare the data for fitting a binomial
> y1 <- RobberyConvict$LC.Y
> n1 <- RobberyConvict$LC.N
```

```
> Y <- cbind(y1, n1-y1)
> head(Y, 5)

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

> ### Fit the GLM
> 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)
> 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 Std. Error
(Intercept)                          -0.2568     0.1561
Step.2001                             0.8232     0.0813
I(Feb + Mar + Apr + May + Jun + Jul)  -0.3723     0.1619
I(Aug + Sep + Oct + Nov + Dec)        -0.5007     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 **
---
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

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

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


Call: glarma(y = Y, X = X, type = "Bin", method = "NR", residuals = "Pearson",
    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 coefficients 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 2, 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.

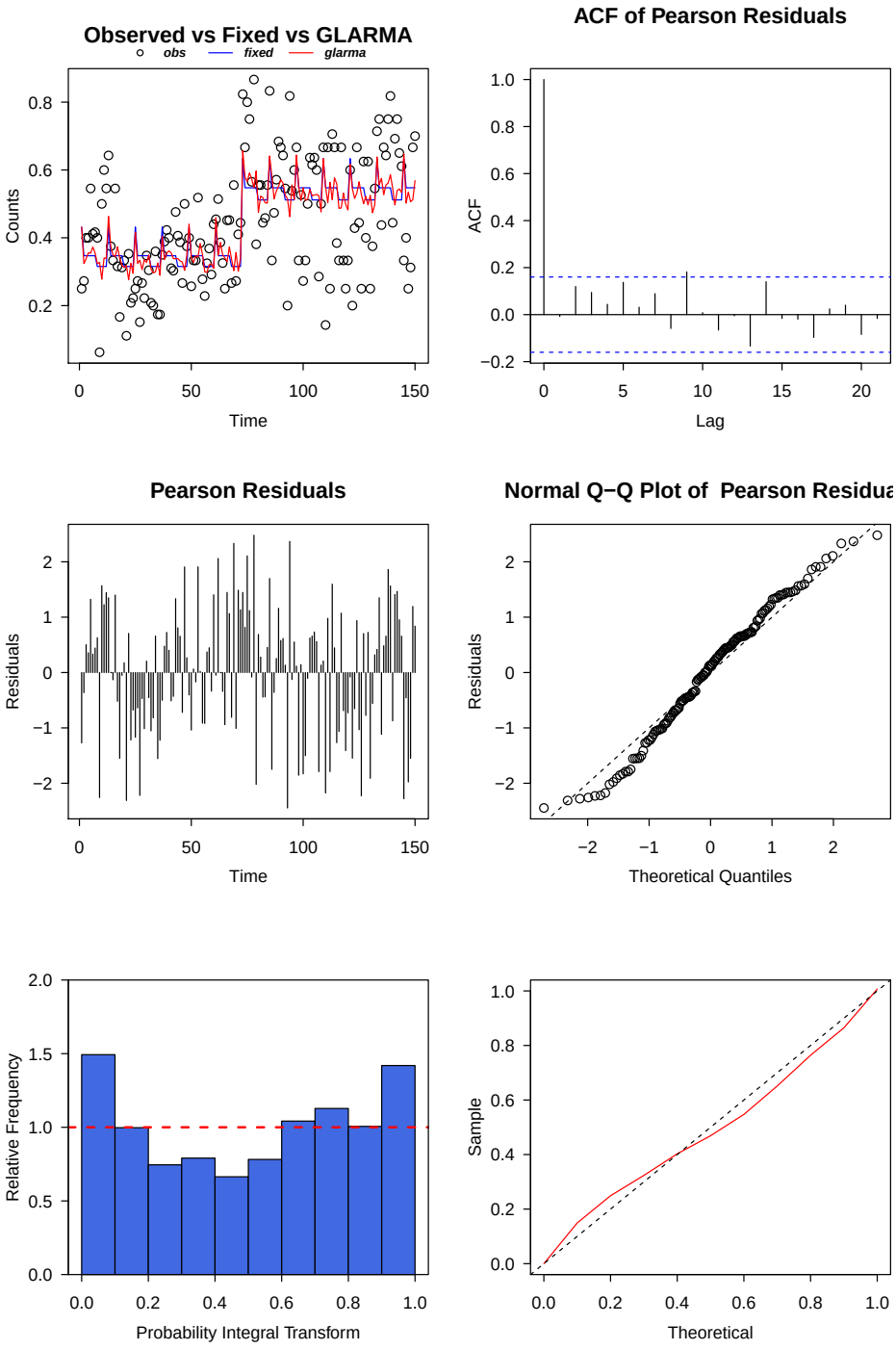

Figure 2: Diagnostic plots for the court conviction model

References

Benjamin MA, Rigby RA, Stasinopoulos DM (2003). "Generalized Autoregressive Moving Average Models." *Journal of the American Statistical Association*, **98**(461), 214–223. doi:10.1198/016214503388619238. URL http://dx.doi.org/10.1198/016214503388619238.

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

Creal D, Koopman SJ, Lucas A (2008). "A General Framework for Observation Driven Time-Varying Parameter Models." *Technical report*, Tinbergen Institute Discussion Paper.

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, Streett SB (2003). "Observation-driven Models for Poisson Counts." *Biometrika*, **90**(4), 777–790.

Davis RA, Dunsmuir WT, Streett SB (2005). "Maximum Likelihood Estimation for an Observation Driven Model for Poisson Counts." *Methodology and Computing in Applied Probability*, **7**(2), 149–159.

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.

Diggle P, Heagerty P, Liang KY, Zeger S (2002). *Analysis of Longitudinal Data*. 2nd edition. Oxford University Press, Oxford.

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.

Hansen BE (1996). "Inference When a Nuisance Parameter is Not Identified under the Null Hypothesis." *Econometrica*, **64**, 413–430.

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

Lu H (2002). *Observation Driven and Parameter Driven Models for Time Series of Counts*. Project report, School of Public Health, University of Minnesota, Minneapolis, MN, USA.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman Hall, London.

Shephard N (1995). "Generalized Linear Autoregressions." *Technical report*, Nuffield College, Oxford University.

Streett S (2000). *Some Observation Driven Models for Time Series of Counts*. Ph.D. thesis, Colorado State University, Department of Statistics, Fort Collins, Colorado.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York, NY. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

Wang B (2004). *GLARMA Models and Stock Price Dynamics*. Project report, School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, 2051, Australia.

Wang C, Li WK (2011). "On the Autopersistence Functions and the Autopersistence Graphs of Binary Autoregressive Time Series." *Journal of Time Series Analysis*, **32**(6), 639–646.

Affiliation:

William T.M. Dunsmuir
Department of Statistics
School of Mathematics and Statistics
University of New South Wales
Sydney, NSW, 2052, Australia
E-mail: W.Dunsmuir@unsw.edu.au
URL: http://web.maths.unsw.edu.au/~dunsmuir/