Censored autoregressive regression models with Student-t innovations

Fernanda Lang Schumacher

Division of Biostatistics - CPH, The Ohio State University

Joint work with: Katherine L. Valeriano, Christian E. Galarza, and Larissa A. Matos

ICSA 2022 Applied Statistics Symposium

Introduction

Introduction

- Observations collected over time are often autocorrelated rather than independent.
- A stochastic model that has been useful in many real-world applications to deal with serial correlation in the error term is the autoregressive (AR) model:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \eta_i.$$

- In general, it is assumed that the disturbance η_t is normally distributed. In practice, however, the normality assumption may be unrealistic, specially in the presence of outliers.
- An additional complication arises when observations are subject to upper or lower detection limits, below and above which they are not quantifiable.

Motivation: Total phosphorus concentration data

- Phosphorus is one of the two nutrients of main concern in Iowa river water, as excessive phosphorus in river water can result in eutrophication.
- Phosphorus concentration data (measured in mg/l) of West Fork Cedar River at Finchford, Iowa, USA, that were collected under the ambient water quality program conducted by the Iowa Department of Natural Resources (Iowa DNR) are available in the R package **carx**.
- The phosphorus concentration measurement was subject to a LOD of 0.02, 0.05, or 0.10, depending on the time, and therefore 15.47% observations are left-censored.
- A gap from 09/2008 to 03/2009 in the data is due to program suspension caused by a temporary lack of funding.

Motivation: Total phosphorus concentration data

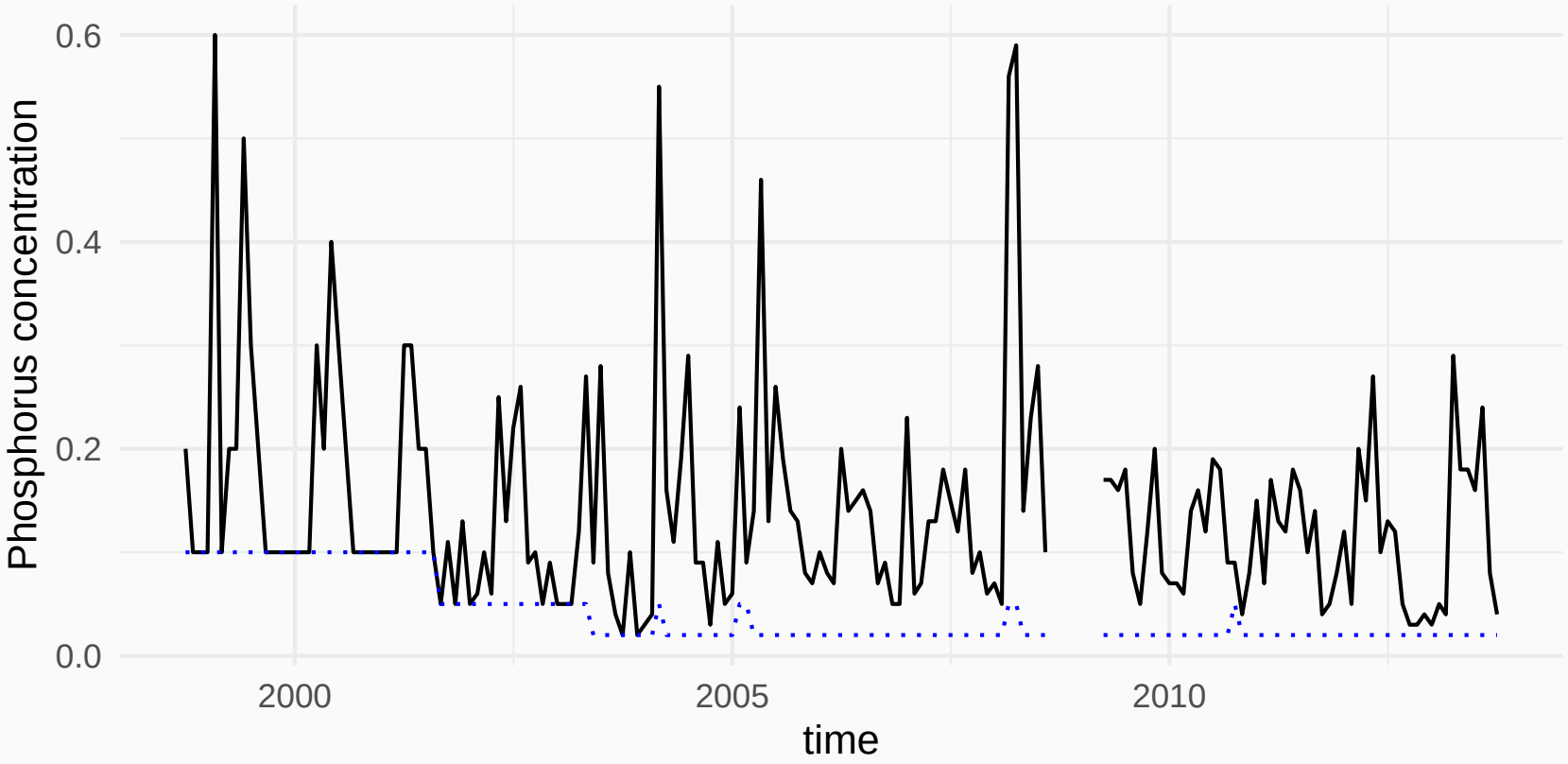


Figure 1: Phosphorus concentration time series and its limit of detection.

Motivation: Total phosphorus concentration data

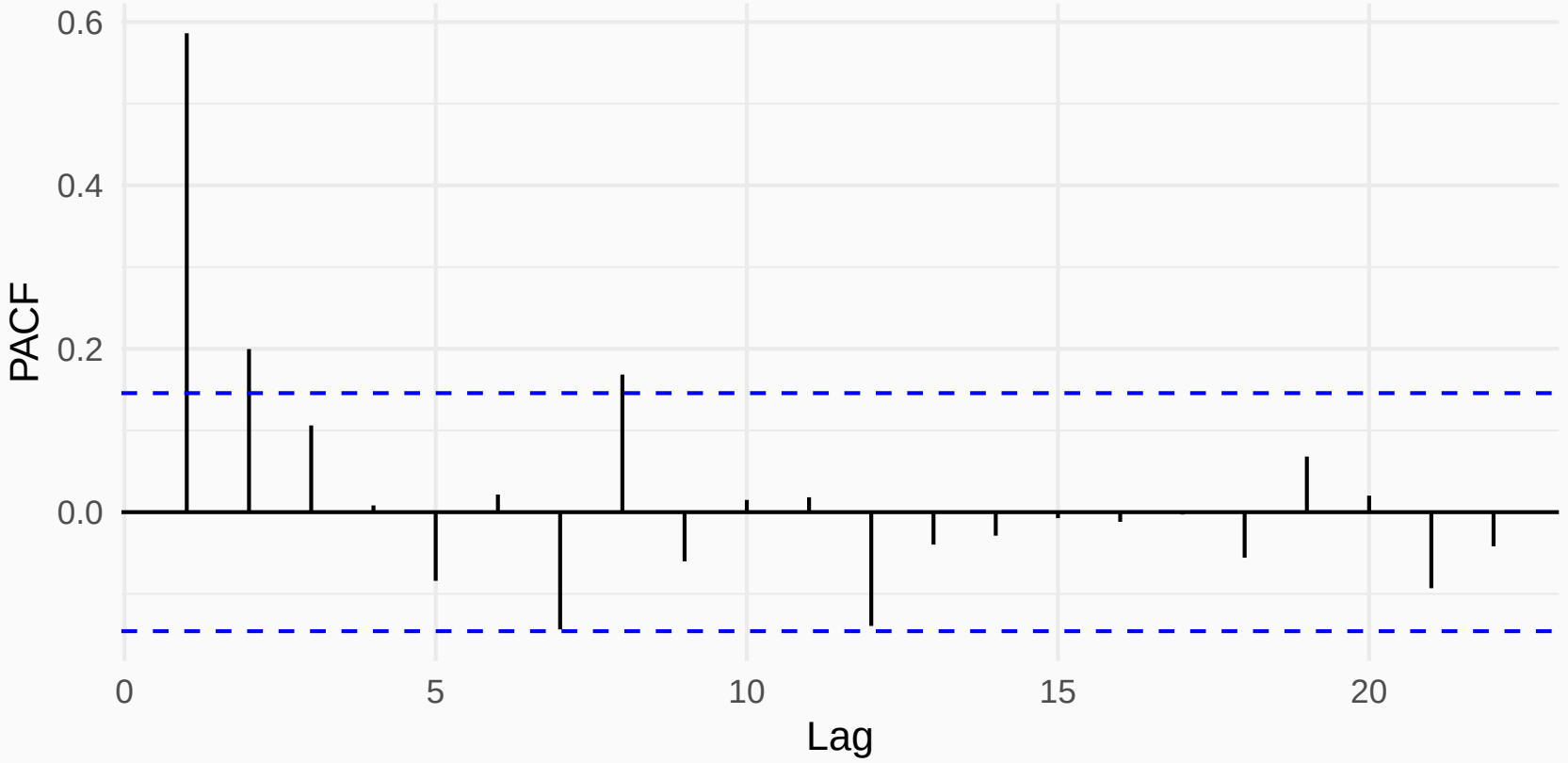


Figure 2: PACF from martingale residuals (Garay et al., 2017) for the censored model with independent errors.

Introduction

- Censored time series data are frequently encountered in environmental monitoring, medicine, and economics.
- Schumacher et al. (2017) considered an autoregressive censored linear model (CAR) estimation via an efficient stochastic approximation of the EM (SAEM) algorithm.
- Liu et al. (2019) proposed a coupled MCMC-SAEM algorithm to fit an AR(p) regression model with Student-t innovations accounting for missing data.

Introduction

- Censored time series data are frequently encountered in environmental monitoring, medicine, and economics.
- Schumacher et al. (2017) considered an autoregressive censored linear model (CAR) estimation via an efficient stochastic approximation of the EM (SAEM) algorithm.
- Liu et al. (2019) proposed a coupled MCMC-SAEM algorithm to fit an AR(p) regression model with Student-t innovations accounting for missing data.
- In this work, we proposed an EM-type algorithm to estimate a CARt(p) model.

Preliminaries

The autoregressive regression model of order p

A linear regression model with errors that are autocorrelated as a discrete-time autoregressive process of order p can be written as

$$\begin{aligned} y_t &= \mathbf{x}_t^\top \boldsymbol{\beta} + \xi_t, \\ \xi_t &= \phi_1 \xi_{t-1} + \ldots + \phi_p \xi_{t-p} + \eta_t, \quad \eta_t \overset{iid}{\sim} F(\cdot), \end{aligned}$$

for $t = 1, \ldots, n$, where

- y_t represents the dependent variable observed at time t,
- $\mathbf{x}_t = (x_{t1}, \ldots, x_{tq})^\top$ is a $q \times 1$ vector of known covariables,
- $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown regression parameters, and
- ξ_t is the regression error and follows an autoregressive model, with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$ being a $p \times 1$ vector of autoregressive coefficients and η_t being a shock of disturbance with distribution $F(\cdot)$.

The autoregressive regression model of order p

- Now, suppose that $\eta_t \sim t(0, \sigma^2, \nu)$. Then, the censored regression model with autoregressive errors will be called the *autoregressive regression t model of order p* (AR$t(p)$).

The autoregressive regression model of order p

- Now, suppose that $\eta_t \sim t(0, \sigma^2, \nu)$. Then, the censored regression model with autoregressive errors will be called the *autoregressive regression t model of order p* (AR$t(p)$).

- The distribution of η_t might be written as $\eta_t = Z_t/\sqrt{U_t}$, with $Z_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \perp U_t \overset{iid}{\sim} \text{Gamma}(\nu/2, \nu/2)$. Consequently, we have the following hierarchical representation:

$$\begin{aligned} \eta_t|(U_t = u_t) &\overset{ind}{\sim} \mathcal{N}(0, \sigma^2/u_t), \\ U_t &\overset{iid}{\sim} \text{Gamma}(\nu/2, \nu/2), \quad t = 1, \ldots, n. \end{aligned}$$

CARt(p) model

The censored AR$t(p)$ model

- Assume that Y_t is not fully observed for all times. Instead, we observe (V_t, C_t), where V_t represents either an observed value or the LOD of a censored variable, and C_t is the censoring indicator defined as

$$C_t = \begin{cases} 1 & \text{if} \quad V_{t1} \leq Y_t \leq V_{t2}, \quad \text{(censored)} \\ 0 & \text{if} \quad V_t = Y_t. \quad \text{(observed)} \end{cases}$$

  The AR$t(p)$ model with censored observations will be called *censored autoregressive regression t model of order p* (CAR$t(p)$).

The censored AR$t(p)$ model

- Assume that Y_t is not fully observed for all times. Instead, we observe (V_t, C_t), where V_t represents either an observed value or the LOD of a censored variable, and C_t is the censoring indicator defined as

$$C_t = \begin{cases} 1 & \text{if} \quad V_{t1} \leq Y_t \leq V_{t2}, \quad \text{(censored)} \\ 0 & \text{if} \quad V_t = Y_t. \quad \text{(observed)} \end{cases}$$

  The AR$t(p)$ model with censored observations will be called *censored autoregressive regression t model of order p* (CAR$t(p)$).

- Left-censored if $C_t = 1$ and $V_t = (-\infty, V_{t2}]$;
- Right-censored if $C_t = 1$ and $V_t = [V_{t1}, +\infty)$;
- Interval censored if $C_t = 1$ and $V_t = [V_{t1}, V_{t2}]$; and
- Missing value if $C_t = 1$ and $V_t = (-\infty, +\infty)$.

The log-likelihood function

- Let $\mathbf{y}_{(t,p)} = (y_{t-1}, \ldots, y_{t-p})^\top$, then we have

$$Y_t|\boldsymbol{\theta}, \mathcal{F}_{t-1} \stackrel{d}{=} Y_t|\boldsymbol{\theta}, \mathbf{y}_{(t,p)} \sim t(\mu_t, \sigma^2, \nu),$$

  where $\mu_t = \mathbf{x}_t^\top \boldsymbol{\beta} + (\mathbf{y}_{(t,p)} - \mathbf{X}_{(t,p)}\boldsymbol{\beta})^\top \boldsymbol{\phi}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top, \sigma^2, \nu)^\top$.

- Conditionally on the first p observations, the observed log-likelihood function can be computed by

$$\ell(\boldsymbol{\theta}; \mathbf{y}_o) = \log \left(\int_{\mathbb{R}^m} \prod_{t=p+1}^{n} f(y_t|\boldsymbol{\theta}, y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \mathrm{d}\mathbf{y}_m \right).$$

Parameter estimation an EM-type algorithm

- On the other hand, letting $\mathbf{y} = (y_{p+1}, \ldots, y_n)^\top$ and $\mathbf{u} = (u_{p+1}, \ldots, u_n)^\top$ be hypothetical missing data, we have $\mathbf{y}_c = \left(\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{V}^\top, \mathbf{C}^\top\right)^\top$, and the complete log-likelihood function can be written as

$$\ell_c(\boldsymbol{\theta}; \mathbf{y}_c) = g(\nu|\mathbf{u}) - \frac{n-p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=p+1}^{n} u_i \left(\tilde{y}_i - \mathbf{w}_i^\top \boldsymbol{\phi}\right)^2 + \text{ cte},$$

with
$g(\nu|\mathbf{u}) = \frac{n-p}{2}\left[\nu \log\left(\frac{\nu}{2}\right) - 2\log\Gamma\left(\frac{\nu}{2}\right)\right] + \frac{\nu}{2}\left(\sum_{i=p+1}^{n} \log u_i - \sum_{i=p+1}^{n} u_i\right)$,
$\tilde{y}_t = y_t - \mathbf{x}_t^\top \boldsymbol{\beta}$, and $\mathbf{w}_t = \mathbf{y}_{(t,p)} - \mathbf{X}_{(t,p)}\boldsymbol{\beta}$.

Parameter estimation an EM-type algorithm

- On the other hand, letting $\mathbf{y} = (y_{p+1}, \ldots, y_n)^\top$ and $\mathbf{u} = (u_{p+1}, \ldots, u_n)^\top$ be hypothetical missing data, we have $\mathbf{y}_c = \left(\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{V}^\top, \mathbf{C}^\top\right)^\top$, and the complete log-likelihood function can be written as

$$\ell_c(\boldsymbol{\theta}; \mathbf{y}_c) = g(\nu|\mathbf{u}) - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=p+1}^{n} u_i\left(\tilde{y}_i - \mathbf{w}_i^\top\boldsymbol{\phi}\right)^2 + \text{ cte},$$

with
$g(\nu|\mathbf{u}) = \frac{n-p}{2}\left[\nu\log\left(\frac{\nu}{2}\right) - 2\log\Gamma\left(\frac{\nu}{2}\right)\right] + \frac{\nu}{2}\left(\sum_{i=p+1}^{n}\log u_i - \sum_{i=p+1}^{n} u_i\right)$,
$\tilde{y}_t = y_t - \mathbf{x}_t^\top\boldsymbol{\beta}$, and $\mathbf{w}_t = \mathbf{y}_{(t,p)} - \mathbf{X}_{(t,p)}\boldsymbol{\beta}$.

- Therefore, for the E-step we need $\widehat{u}^{(k)} = \sum_{i=p+1}^{n}\widehat{u}_i^{(k)}$,
$\widehat{u}_i^{(k)} = \mathbb{E}[U_i|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{\log(\mathbf{u})}^{(k)} = \mathbb{E}[\sum_{i=p+1}^{n}\log(U_i)|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$,
$\widehat{uy^2}^{(k)} = \mathbb{E}[\sum_{i=p+1}^{n} U_i Y_i^2|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{uy_i}^{(k)} = \mathbb{E}[U_i Y_i|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$,
$\widehat{uy\mathbf{y}}^{(k)} = \mathbb{E}[\sum_{i=p+1}^{n} U_i Y_i \mathbf{Y}_{(i,p)}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{u\mathbf{y}_i}^{(k)} = \mathbb{E}[U_i\mathbf{Y}_{(i,p)}|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$, and
$\widehat{u\mathbf{y}^2}^{(k)} = \mathbb{E}[\sum_{i=p+1}^{n} U_i\mathbf{Y}_{(i,p)}\mathbf{Y}_{(i,p)}^\top|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k-1)}]$, for $i = p+1, \ldots, n$.

The SAEM algorithm

Delyon et al. (1999) proposed splitting the E-step from the EM algorithm into a simulation step and an integration step (using a stochastic approximation):

- **E-step**:
  1. Simulation: Draw M samples of the missing data $\mathbf{y}_m^{(l,k)}, l = 1, \ldots, M$ from the conditional distribution $f(\mathbf{y}_m; \widehat{\boldsymbol{\theta}}^{(k)}, \mathbf{y}_o)$.
  2. Stochastic approximation: Update $Q_k(\boldsymbol{\theta})$ by

$$\widehat{Q}_k(\boldsymbol{\theta}) = \widehat{Q}_{k-1}(\boldsymbol{\theta}) + \delta_k \left(\frac{1}{M} \sum_{l=1}^{M} \ell_c(\boldsymbol{\theta}; \mathbf{y}_m^{(k,l)}, \mathbf{y}_o) - \widehat{Q}_{k-1}(\boldsymbol{\theta}) \right),$$

  where δ_k is a decreasing sequence of positive numbers such that $\sum_{k=1}^{\infty} \delta_k = \infty$ and $\sum_{k=1}^{\infty} \delta_k^2 < \infty$, also known as the smoothness parameter (Kuhn & Lavielle, 2005).

Prediction

- Let $\mathbf{y}_{\text{obs}} = (\mathbf{y}_o^\top, \mathbf{y}_m^\top)^\top$, where $\mathbf{y}_o$ is the vector of uncensored observations and $\mathbf{y}_m$ is the vector of censored or missing observations.
- To deal with the censored values in $\mathbf{y}_{\text{obs}}$, we use an imputation procedure that consists in replacing the censored observations with the values obtained in the last iteration of the SAEM algorithm

$$\mathbf{y}_m = \mathbb{E}[\mathbf{y}_m|\mathbf{V}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(W)}] \approx \widehat{\mathbf{y}}_m^{(W)}.$$

- The imputed vector will be denoted by $\mathbf{y}_{\text{obs}*} = (\mathbf{y}_o^\top, \widehat{\mathbf{y}}_m^{(W)\top})^\top$.
- Now, the forecasting procedure will be performed recursively as follows:

$$\widehat{y}_{n+k} = \begin{cases} \mathbf{x}_{n+k}^\top \boldsymbol{\beta} + \sum_{j=k}^{p} \phi_j (y_{n+k-j} - \mathbf{x}_{n+k-j}^\top \boldsymbol{\beta}), & k = 1 \\ \mathbf{x}_{n+k}^\top \boldsymbol{\beta} + \sum_{i=1}^{k-1} \phi_i (\widehat{y}_{n+k-i} - \mathbf{x}_{n+k-i}^\top \boldsymbol{\beta}) + & \\ \quad \sum_{j=k}^{p} \phi_j (y_{n+k-j} - \mathbf{x}_{n+k-j}^\top \boldsymbol{\beta}), & 1 < k \leq p \\ \mathbf{x}_{n+k}^\top \boldsymbol{\beta} + \sum_{j=1}^{p} \phi_j (\widehat{y}_{n+k-j} - \mathbf{x}_{n+k-j}^\top \boldsymbol{\beta}), & p < k \leq n_{\text{pred}}. \end{cases}$$

Residual analysis

- To perform residual analysis for the CAR$t(p)$ model, we first impute the censored or missing observations.
- Then, we consider quantile residuals computed as

$$\widehat{r}_i = \Phi^{-1}\left(T_1\left(y_i; \widehat{\mu}_i, \widehat{\sigma}^2, \widehat{\nu}\right)\right), \quad i = p+1, \ldots n,$$

  where $\widehat{\mu}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + (\mathbf{y}_{(i,p)} - \mathbf{X}_{(i,p)}\widehat{\boldsymbol{\beta}})^\top \widehat{\boldsymbol{\phi}}$ and$\widehat{\boldsymbol{\theta}}$ refers to the ML estimates of $\boldsymbol{\theta}$ obtained through the SAEM algorithm.
- Note that the quantile residual is calculated only for the latest $n - p$ observations.

Application

Application: Phosphorus concentration data

- Aiming to evaluate the prediction accuracy, the dataset was train-test split. The training dataset consists of 169 observations, where 20.71% are left-censored or missing, while the testing dataset contains 12 observations, all fully observed.

- Following Wang & Chan (2018), to make the phosphorous concentration (P) and the water discharge (Q) relation more linear we considered the logarithmic transformation of P and Q and fitted the following model:

$$\log(P_t) = \sum_{j=1}^{4} [\beta_{0j} S_{jt} + \beta_{1j} S_{jt} \log(Q_t)] + \xi_t, \quad t = p+1, \ldots, n,$$

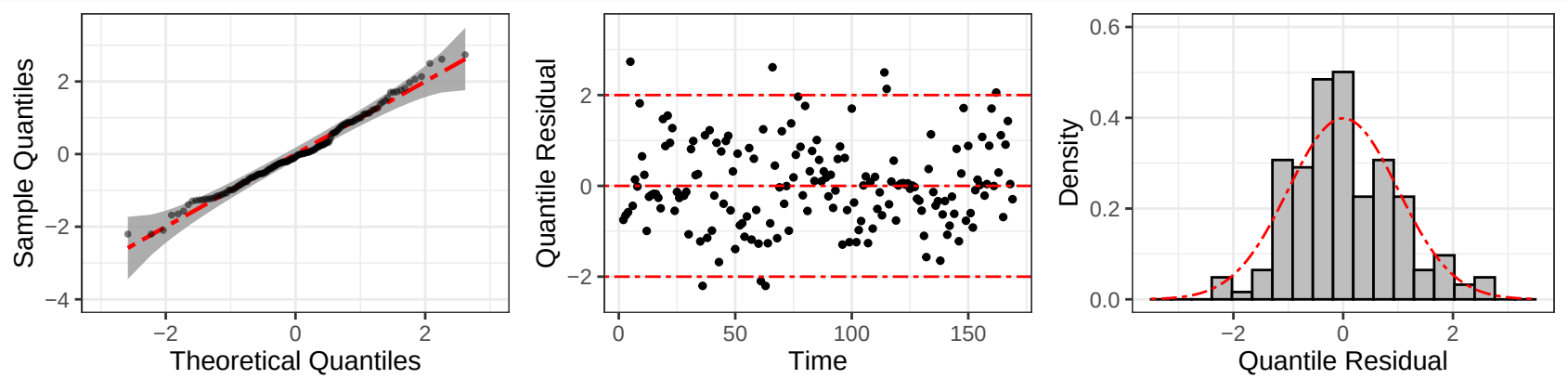
  where the regression error ξ_t follows an autoregressive model and S_j is a dummy seasonal variable for quarters $j = 1, 2, 3$, and 4.

Application: ML estimation

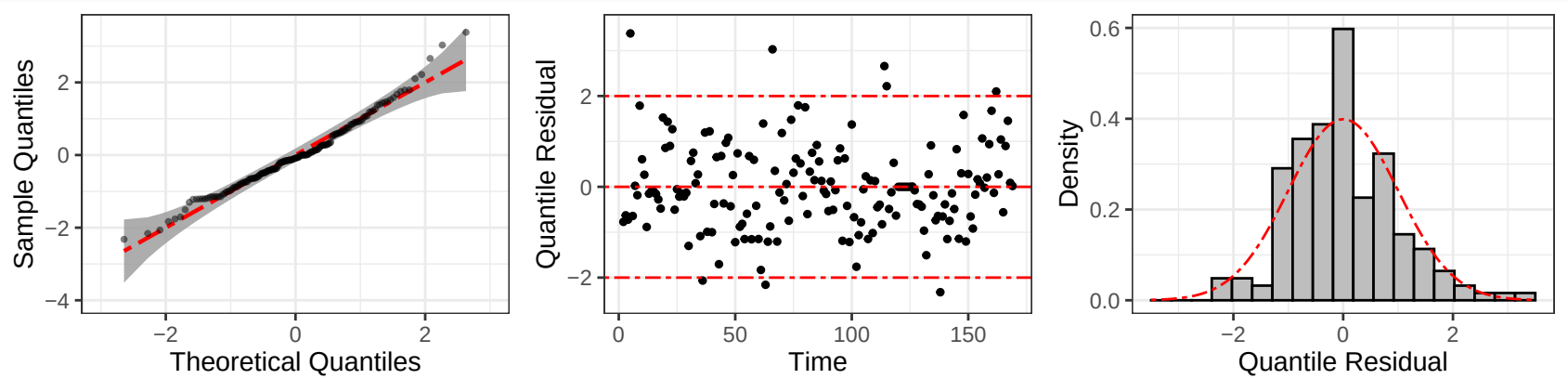
Parameters	CARt(1)			CAR(1)		
	Estimate	SE	95% C.I.	Estimate	SE	95% C.I.
β_{01}	-4.338	0.609	(-5.532 ; -3.144)	-4.670	0.513	(-5.675 ; -3.665)
β_{02}	-2.731	0.750	(-4.201 ; -1.261)	-3.022	0.751	(-4.494 ; -1.550)
β_{03}	-4.141	0.419	(-4.962 ; -3.320)	-4.174	0.442	(-5.040 ; -3.308)
β_{04}	-4.651	0.545	(-5.719 ; -3.583)	-4.999	0.549	(-6.075 ; -3.923)
β_{11}	0.293	0.107	(0.083 ; 0.503)	0.373	0.085	(0.206 ; 0.540)
β_{12}	0.139	0.105	(-0.067 ; 0.345)	0.185	0.105	(-0.021 ; 0.391)
β_{13}	0.363	0.070	(0.226 ; 0.500)	0.364	0.075	(0.217 ; 0.511)
β_{14}	0.351	0.097	(0.161 ; 0.541)	0.410	0.098	(0.218 ; 0.602)
ϕ_1	-0.087	0.085		-0.077	0.089	
σ^2	0.176	0.046		0.254	0.030	
ν	5.002	3.059		–	–	
MSPE		0.101			0.126	
MAPE		0.240			0.255	

Application: Residual analysis

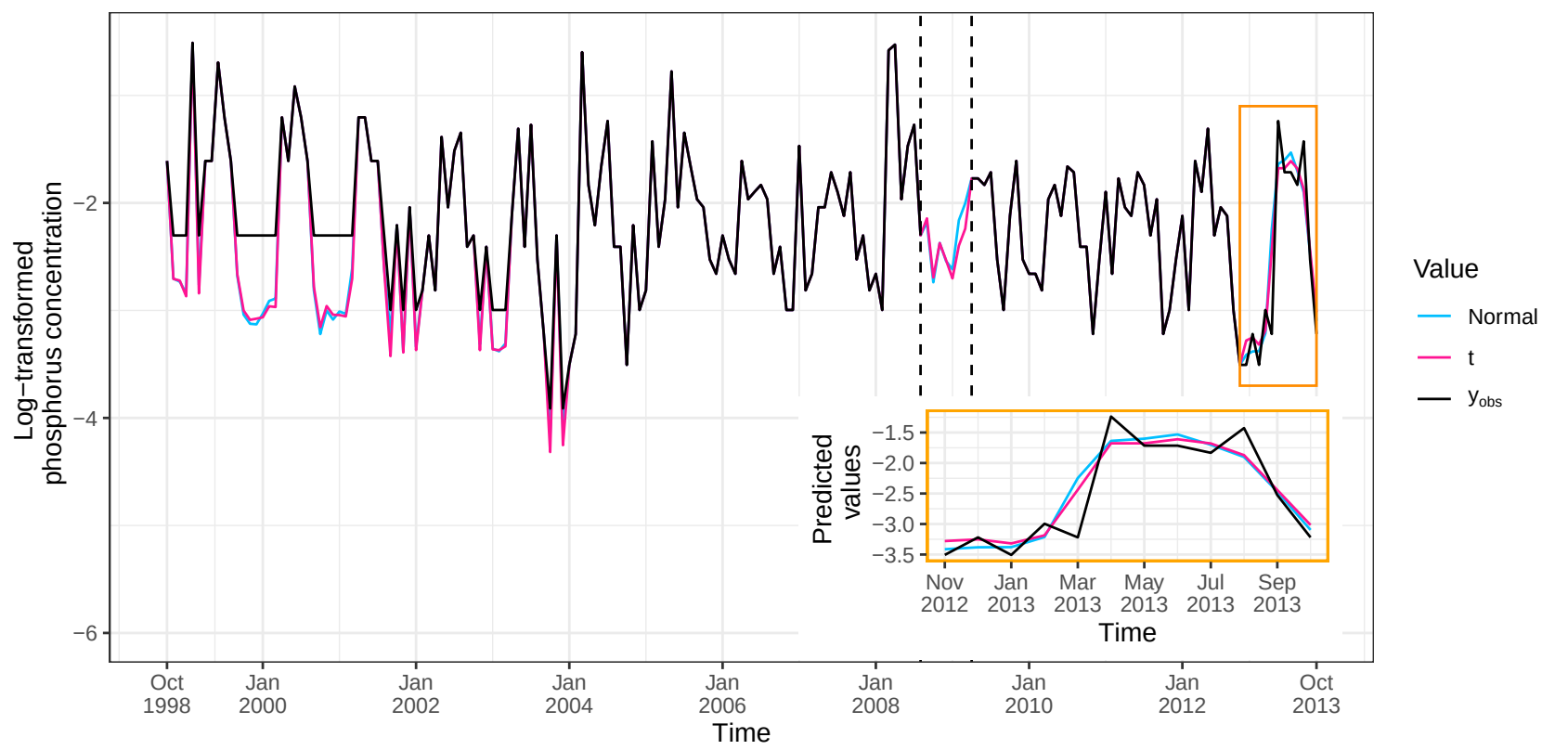
CARt(1)



CAR(1)



Application: Model fit for phosphorus concentration data



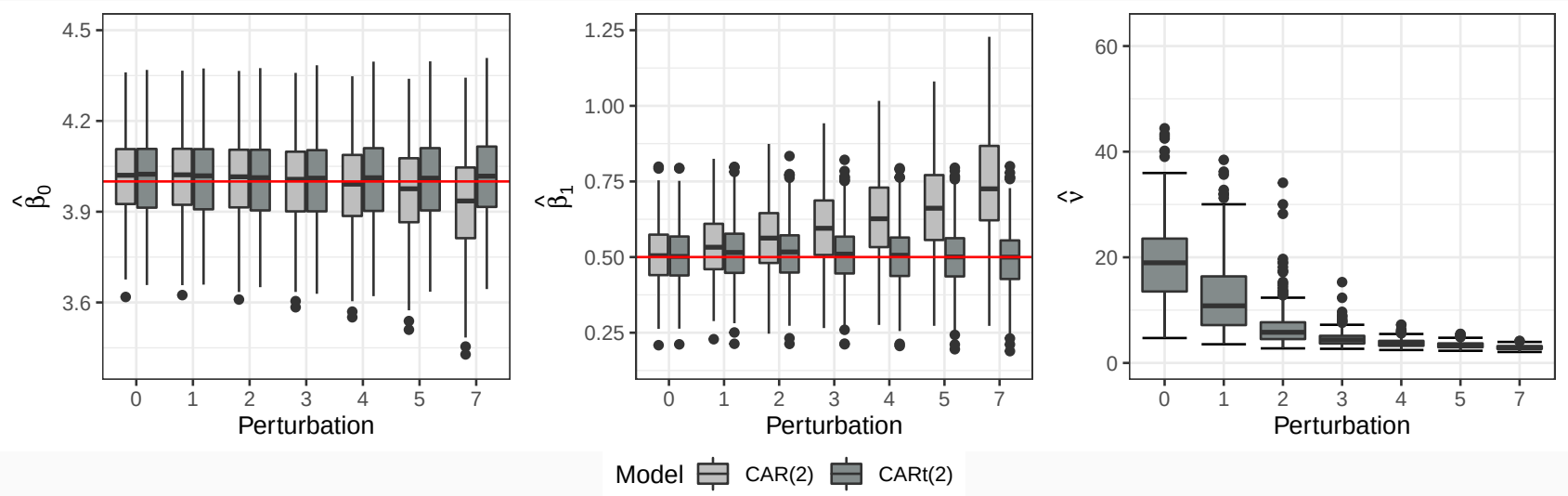
Simulation study

Simulation study: Robustness of the estimators

- Aim: to compare the performance of the estimates for two censored AR models in the presence of outliers on the response variable.
- We generated 300 MC samples of size $n = 100$ under the CAR(2) model (with normal innovations).
- After generating the data, each MC sample was perturbed as $\max(\mathbf{y}) = \max(\mathbf{y}) + \vartheta \mathrm{SD}(\mathbf{y})$, for $\vartheta \in \{0, 1, 2, 3, 4, 5, 7\}$.

Simulation study: Robustness of the estimators

Level of censoring: 30%



Concluding remarks

Concluding remarks

- This work introduces a novel model that can handle left, right, or interval censoring time series, while simultaneously modeling heavy-tails and missing observations.
- The methods developed in this work are implemented at the R package **ARCensReg**, whose current version is available at GitHub.
- It is important to remark that we assumed the dropout/censoring mechanism to be missing at random (MAR). A possible venue for future work is to consider scenarios with informative censoring.
- Another interesting extension is to tackle the limitation of assuming that the first p observations are fully observed.

References

- Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1), 94–128.
- Garay, A., Lachos, V.H., Bolfarine, H., & Cabral, C. (2017). Linear Censored Regression Models with Scale Mixture of Normal Distributions. *Statistical Papers*, 58:247–278.
- Kuhn, E. & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*, 49(4), 1020–1038.
- Liu, J., Kumar, S., & Palomar, D.P. (2019). Parameter estimation of heavy-tailed AR model with missing data via stochastic EM. *IEEE Transactions on Signal Processing*, 67(8), 2159–2172.
- Schumacher, F.L., Lachos, V.H., & Dey, D.K. (2017). Censored regression models with autoregressive errors: A likelihood-based perspective. *Canadian Journal of Statistics*, 45(4), 375–392.
- Schumacher, F.L., Valeriano, K., Lachos, V.H., Galarza, C.E., & Matos, L.A. (2022). Package 'ARCensReg'. R package version 3.0.0.
- Wang, C. & Chan, K.-S. (2018). Quasi-likelihood estimation of a censored autoregressive model with exogenous variables. *Journal of the American Statistical Association*, 113(523), 1135–1145.

GitHub repository

Fernanda Lang Schumacher
Assistant Professor - Biostatistics
CPH - The Ohio State University
schumacher.313@osu.edu