Supplemental Materials for "Bayesian Negative Binomial Regression Model With Unobserved Covariates for Predicting the Frequency of North Atlantic Tropical Storms"

Xun Li, Joyee Ghosh, and Gabriele Villarini

1 Simulation Results When the Data Generating Model is Negative Binomial

We generate datasets of 100 observations. For each of the simulated datasets, we generate the covariates from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{X} \sim MVN\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \sigma_{(p-1)p} \\ \sigma_{1p} & \cdots & \sigma_{(p-1)p} & \sigma_p^2 \end{pmatrix}\right),$$

where $p = 12$, ${\sigma_i}^2 = 1$, for $i = 1, \ldots, 12$, and the off-diagonal elements of $\boldsymbol{\Sigma}$ are set as $\sigma_{12} = 0.7$, $\sigma_{13} = 0.9$, $\sigma_{14} = 0.7$, $\sigma_{23} = 0.6$, $\sigma_{24} = 0.7$, $\sigma_{34} = 0.7$, and $\sigma_{ij} = 0.5$ elsewhere.

Next we generate an outcome Y_i, using a negative binomial (NB) regression model with mean $\exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right)$ and dispersion parameter $\eta = 1$, where $\boldsymbol{\beta} = (4.5, 1.0, -0.5, 0.875, -0.625, 0, \ldots, 0)$. We split the 100 observations into two equal halves. We use 50 observations for estimation, and 50 observations for prediction, to compare methods. For prediction, we treat $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as missing.

Our goal is to compare results from a NB regression model to a Poisson regression model, when data are generated from a NB regression model with overdispersion. The results using both regression models and different approaches for dealing with the missing covariates (see Section 4.3), averaged over 100 datasets, are presented in Table 1. In order to approximate a Poisson regression model, we fix η in the NB model, at a very large value so that the mean and variance would be about the same, and the model will not allow for overdispersion. We fix η at 10250^3, where 10250 is the maximum value of the mean response under the NB data generation. The results in Table 1 show that the NB regression model performs much better than the Poisson regression model, when there is overdispersion in the data. The NB model has lower RMSE and much better frequentist coverage. The Poisson model has undercoverage, as expected, as it cannot account for the overdispersion in the data.

Method	Cor.Pearson	Cor.Spearman	RMSE	MAE	Coverage		Size	
					Equal-tailed	HPD	Equal-tailed	HPD
No missing (NB)	0.59	0.63	321.04	137.63	0.90	0.89	692.19	489.93
No missing (Poisson)	0.50	0.58	583.22	212.31	0.16	0.16	65.78	65.00
Method 1 (NB)	0.54	0.59	324.78	140.83	0.91	0.89	744.18	496.31
Method 1 (Poisson)	0.46	0.53	494.91	191.39	0.49	0.53	454.23	363.17
Method 2 (NB)	0.54	0.59	323.56	141.86	0.90	0.89	682.89	479.11
Method 2 (Poisson)	0.46	0.53	578.39	213.69	0.14	0.14	58.88	58.09

Table 1: Results related to the predictive distribution of the response variable, when data are generated from the negative binomial regression model with overdispersion. Method 1 retains all covariates; Method 2 discards the covariates with missing values. Results are averaged over 100 simulated datasets.

2 Additional Results for the Tropical Storm Example

In the main manuscript we have demonstrated that the NB regression model works well for our dataset. NB regression was mainly adopted for computational convenience. Because the dataset does not have overdispersion, we could have alternatively used Poisson regression. We approximate Poisson regression by fixing the overdispersion parameter at 19^3, where 19 is the maximum count of tropical storms in the dataset. The results for July are given in Table 2. The NB and Poisson regression models give very similar results, which is expected because the dataset does not exhibit overdispersion. Slightly smaller prediction sets under Poisson regression is also expected because the Poisson model does not allow for overdispersion.

Method	Cor.Pearson	Cor.Spearman	RMSE	MAE	Coverage		Size	
					Equal-tailed	HPD	Equal-tailed	HPD
Method 1 (NB)	0.76	0.82	1.87	1.25	1.00	1.00	14.13	13.38
Method 1 (Poisson)	0.76	0.82	1.87	1.25	1.00	1.00	13.88	13.38
Method 2 (NB)	0.71	0.70	2.00	1.50	1.00	1.00	13.50	13.13
Method 2 (Poisson)	0.71	0.70	2.00	1.50	1.00	1.00	13.38	13.00

Table 2: Results for July.

3 Diagnostic Plots for the Tropical Storm Example

In this section we first provide the list of response variables and predictors in the models used for July, in the analysis of the tropical storm dataset. Next, we provide the associated diagnostic plots that are available in R for normal linear regression (for modeling the SSTs) and Poisson regression (for modeling the frequency of tropical storms) models. Based on the autocorrelation plots, the assumption of independence of the response variables in the two levels of the model seems reasonable. The assumption of normality for the second level models for SSTs also seems more or less reasonable. The plots for other months are similar so we do not report them here.

Response				
TS	OBS_{Atl}	OBS_{Trop}	$CMC2_{Atl}$	$CMC2_{Trop}$
intercept	intercept	intercept	intercept	intercept
$GFDLA_{Atl}$	$GFDLA_{Atl}$		$GFDLA_{Atl}$	$GFDLA_{Atl}$
$GFDLB_{Atl}$	$GFDLB_{Atl}$		$GFDLB_{Atl}$	$GFDLB_{Atl}$
$GFDL_{Atl}$	$GFDL_{Atl}$		$GFDL_{Atl}$	$GFDL_{Atl}$
$NASA_{Atl}$	$NASA_{Atl}$		$NASA_{Atl}$	$NASA_{Atl}$
$GFDLA_{Trop}$	$GFDLA_{Trop}$		$GFDLA_{Trop}$	$GFDLA_{Trop}$
$GFDLB_{Trop}$	$GFDLB_{Trop}$		$GFDLB_{Trop}$	$GFDLB_{Trop}$
$GFDL_{Trop}$	$GFDL_{Trop}$	$GFDL_{Trop}$	$GFDL_{Trop}$	$GFDL_{Trop}$
$NASA_{Trop}$	$NASA_{Trop}$		$NASA_{Trop}$	$NASA_{Trop}$
OBS_{Atl}		OBS_{Atl}	OBS_{Atl}	
OBS_{Trop}			OBS_{Trop}	OBS_{Trop}
$CMC2_{Atl}$				$CMC2_{Atl}$
$CMC2_{Trop}$				

Table 3: Models for July.

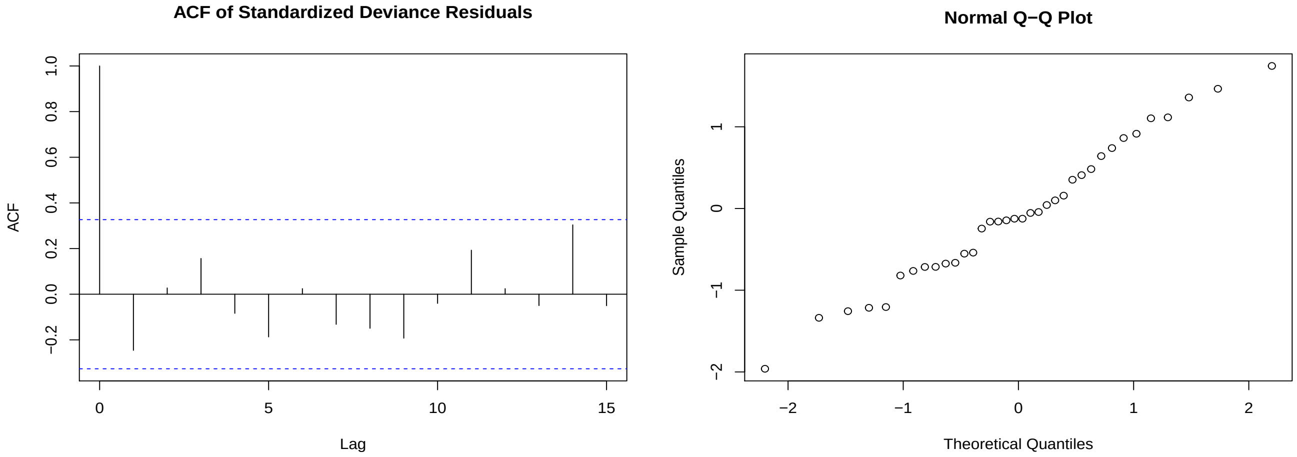

Figure 1: Residual plots for examining independence and normality of TS.

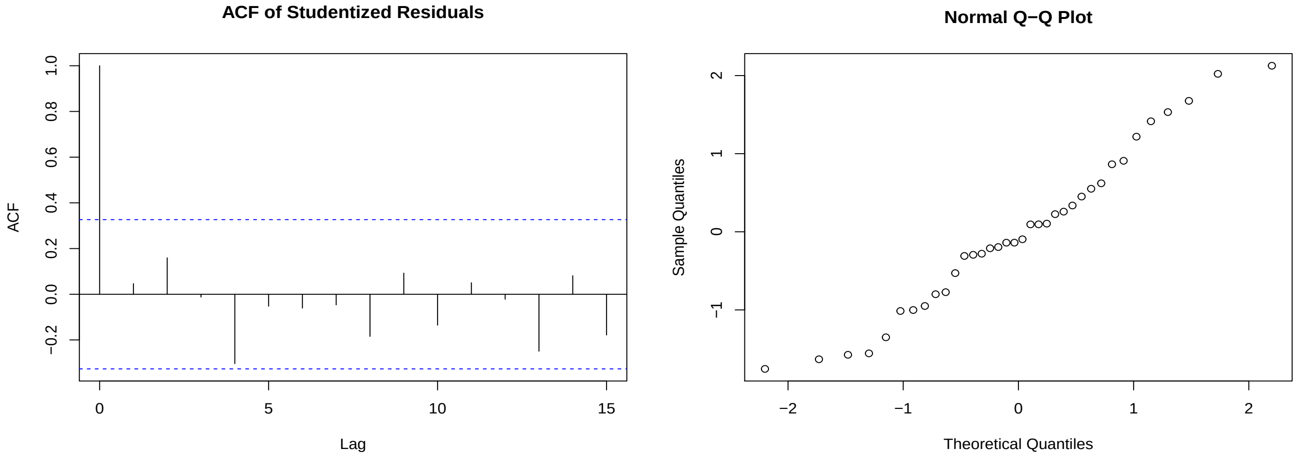

Figure 2: Residual plots for examining independence and normality of Obs_{Atl}.

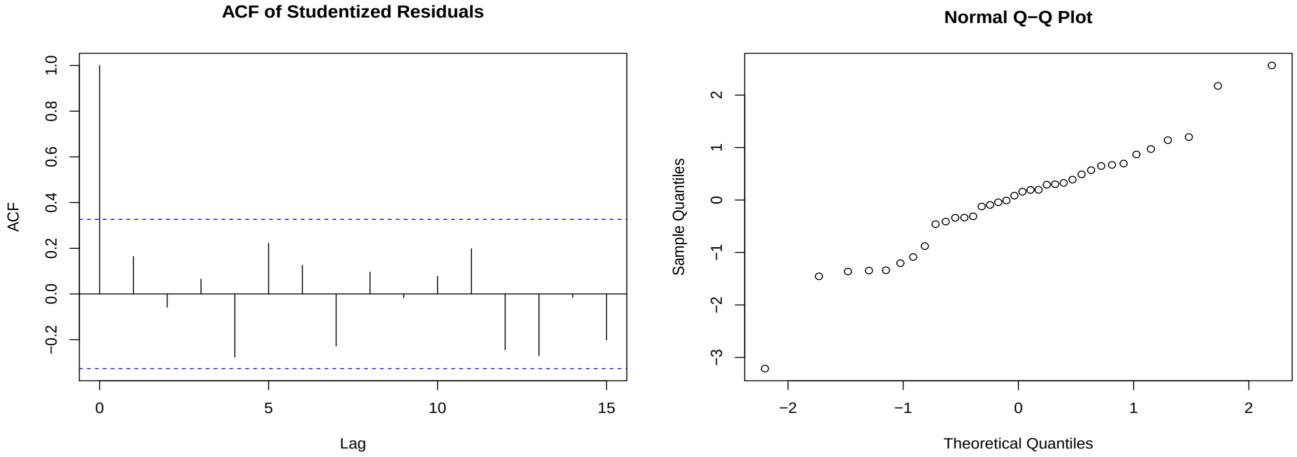

Figure 3: Residual plots for examining independence and normality of Obs_{Trop}.

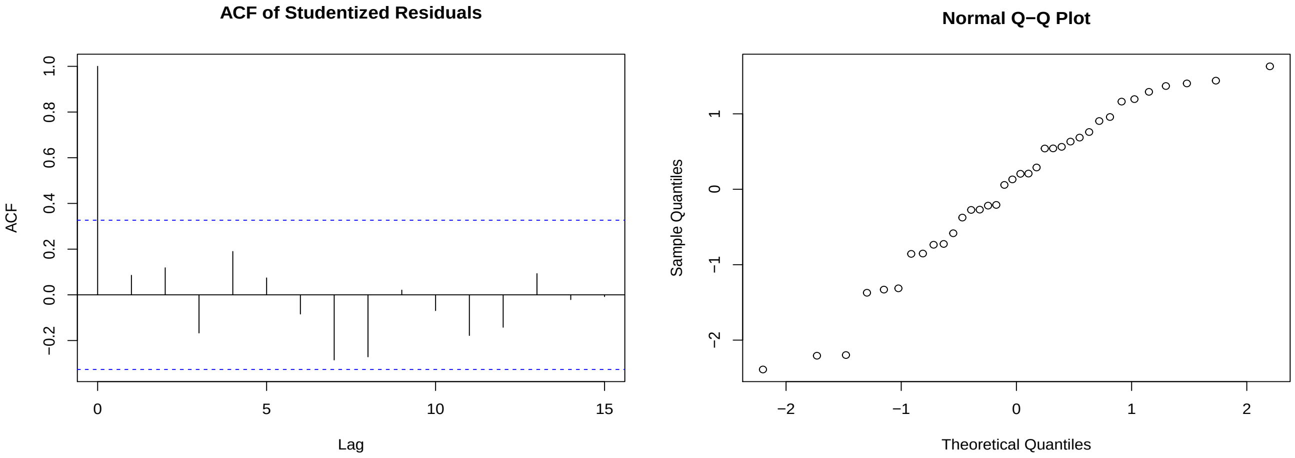

Figure 4: Residual plots for examining independence and normality of $CMC2_{Atl}$.

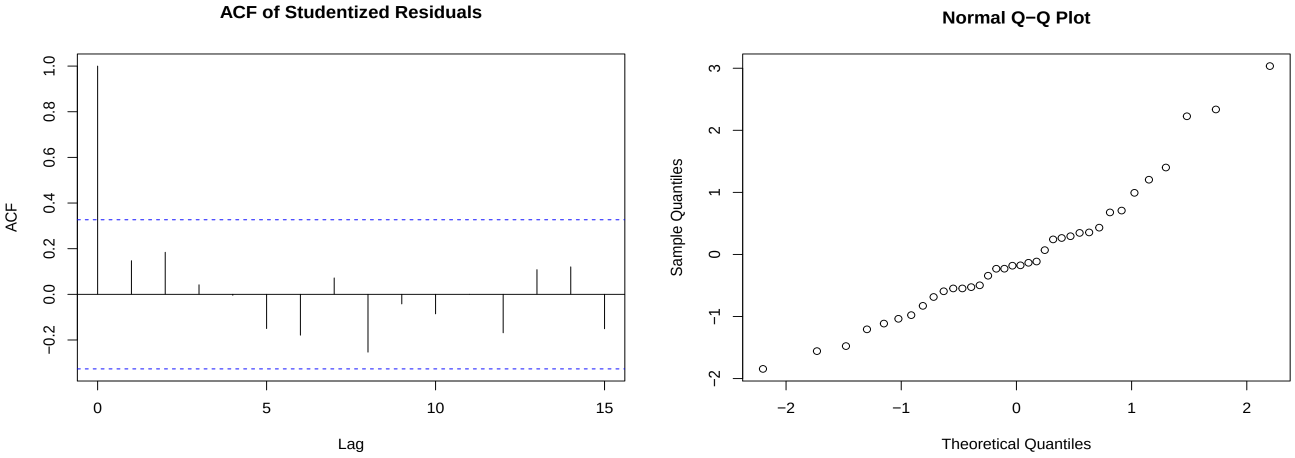

Figure 5: Residual plots for examining independence and normality of $CMC2_{Trop}$.