A Nonlinear Regression Model for Modeling Autocorrelated and Overdispersed Count Data

Examination Roll: 144210

Department of Statistics
University of Dhaka

September, 2024

Outline

- Introduction
- Literature Review
- Motivation
- Methodology
- Simulation Study
- Application
- Conclusion

Autocorrelated and Overdispersed Count Data

- **Count Data:** Derived from counting
  - **Example:** Number of road accidents in a day
- **Time Series Count Data:** Counts obtained over time
  - **Example:** Daily car accidents in a city over the course of a year
- **Autocorrelation:** Temporal dependence $\rightarrow$ Overdispersion

Real Life Example of Autocorrelation

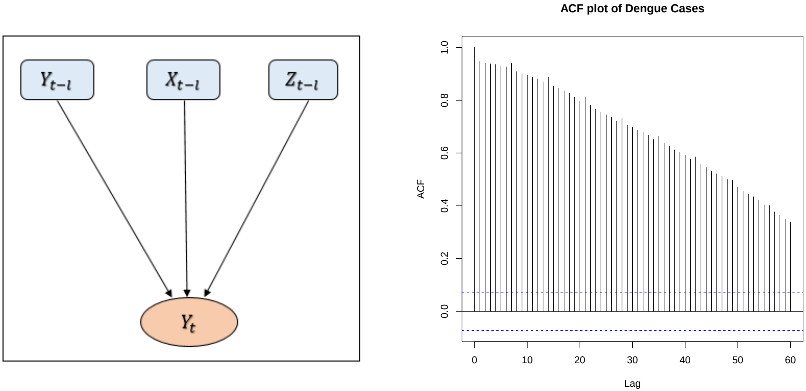

Figure 1: Temporal dependence

- Y = Dengue Cases and covariates (X = Temperature, Z = Humidity)

Source: Directorate General of Health Services (DGHS), 2022-23

Real Life Examples of Nonlinearity (DGHS)

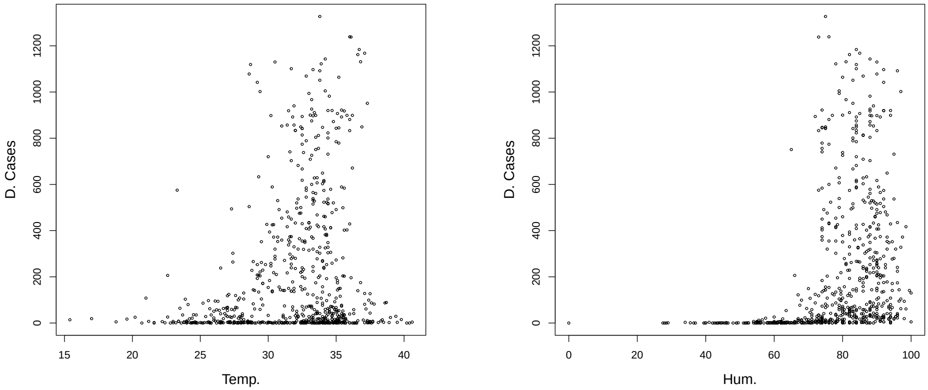

Figure 2: Scatter plots of predictors vs response

Response (Dengue Cases, **Source:** DGHS 2022-23) and Predictors (Temperature and Humidity, **Source:** BMD)

Literature Review

The effects of climate variables on dengue cases were explored by

- **Linear Models**
  - Sharmin et al. (2015): Negative Binomial Generalized Linear Model
  - Islam et al. (2021): Poisson, Zero-Inflated Poisson and Negative Binomial
- **Nonlinear Models**
  - Islam et al. (2023): Poisson Generalized Additive Model
  - Hossain et al. (2023): Quasi-Poisson and Zero-Inflated Poisson
- **Nonlinearity and Autocorrelation**
  - Generalized Additive Model with Autocorrelation (GAMAR) proposed by Lei et al. (2012)

Motivation

Limitations:

- **Linear Models:** Failed to incorporate
  - Nonlinearity and temporal dependence
- **Nonlinear Models:** Incorporated nonlinearity but not
  - Temporal dependence
- **GAMAR:** Covered nonlinearity and dependency but not explored
  - Performance of GAMAR under different lags and functional forms

Objectives:

- Explore how GAMAR performs under different sample sizes, lags and functional forms
- Compare the performance of GAM and GAMAR
- Show an application of GAMAR to real life data

Generalized Additive Model with Autoregressive Terms

- **Generalized Additive Model (GAM)**

  Trevor et al. (1990) developed GAM by extending GLM framework as:

$$\ln(\mu_i) = \beta_0 + \underbrace{\sum_{j=1}^{k} f_j(x_{ij})}_{\text{smooth functions}}$$

$$\ln(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik}) .$$

- **Generalized Additive Model with Autoregressive Terms (GAMAR)**

  Lei et al. (2012) developed GAMAR by introducing AR terms in GAM as:

$$\ln(\mu_t) = \underbrace{\sum_{i=1}^{m} \mathrm{ns}(x_{it}, df_i)}_{\text{smoother}} + \underbrace{\sum_{j=1}^{p} c_j \left[\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} \mathrm{ns}\left(x_{(t-j),i}, df_i\right) \right]}_{\text{autoregressive terms } (a_t)},$$

  where $y^*_t = \max(y_t, \tau)$, τ is a positive threshold parameter.

- **Estimation:** Maximum Partial Likelihood

Simulation Study

- **Simulation Setup**

Table 1: Different sample sizes and different lags (AR order)

Sample Size (days)	**AR Order** 1	2	3	4
730	(1, 730)	(2, 730)	(3, 730)	(4, 730)
1461	(1, 1461)	(2, 1461)	(3, 1461)	(4, 1461)
2191	(1, 2191)	(2, 2191)	(3, 2191)	(4, 2191)
2922	(1, 2922)	(2, 2922)	(3, 2922)	(4, 2922)

- $(1, 730) \rightarrow$ Sample size of 730 days, AR order 1
- 16 possible combinations (cases)

Algorithm of Data Generation (Lag 4)

Executing the following steps yield a single observation from the GAMAR (4):

Step 1: Set the true values $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) =$ (5.02, -0.45, -0.46, -0.48, -0.43, -0.38, -0.25) and $(c_1, c_2, c_3, c_4) = (0.5, 0.25, 0.12, 0.06)$

Step 2: Choose a_t from $a_t \sim \text{Normal}\,(4, 0, 0.2)$

Step 3: Simulate $y_t \sim \text{Poisson}\,(\mu_t)$ where

- $\ln(\mu_t) = \sum_{i=1}^{6} \beta_i s_{i6}(x_t) + \sum_{i=1}^{p} c_i \Big(\ln\left(y^*_{t-i}\right) - ns\left(x_{t-i}, 6\right) \Big)$
- $y^*_t = \max(y_t, \tau)$, τ=0.5 and x_t = covariate

Results from GAM and GAMAR (4)

Table 2: Results from GAM and GAMAR (4) ($n = 730$, lag $= 4$)

		GAM				GAMAR (4)			
	TruPar	MeaEst	Bias	RelErr	Coverage	MeaEst	Bias	RelErr	Coverage
β_0	**5.02**	4.9613	-0.0587	0.0237	53.4	5.0050	-0.0150	0.0161	90.7
β_1	**-0.45**	-0.4520	-0.0020	0.1824	66.9	-0.4527	-0.0027	0.0785	95.0
β_2	**-0.46**	-0.4562	-0.0038	0.2254	66.5	-0.4571	-0.0029	0.0976	95.9
β_3	**-0.48**	-0.4789	-0.0011	0.1878	68.8	-0.4811	-0.0011	0.0852	94.2
β_4	**-0.43**	-0.4231	0.0069	0.1185	83.6	-0.4250	0.0050	0.0744	95.0
β_5	**-0.38**	-0.3751	0.0049	0.5383	68.6	-0.3804	-0.0004	0.2417	95.4
β_6	**-0.25**	-0.2519	-0.0019	0.2963	75.6	-0.2495	0.0005	0.1640	94.5
					Mean				
			0.0113	0.2246	69.05		0.0039	0.1082	94.38
c_1	**0.5**					0.4965	-0.0035	0.0981	94.4
c_2	**0.25**					0.2416	-0.0084	0.2188	94.8
c_3	**0.12**					0.1140	-0.0060	0.4635	94.6
c_4	**0.06**					0.0536	-0.0064	0.8114	96.3
					Mean				
							0.0061	0.3979	95.02

ACF and PACF Plots

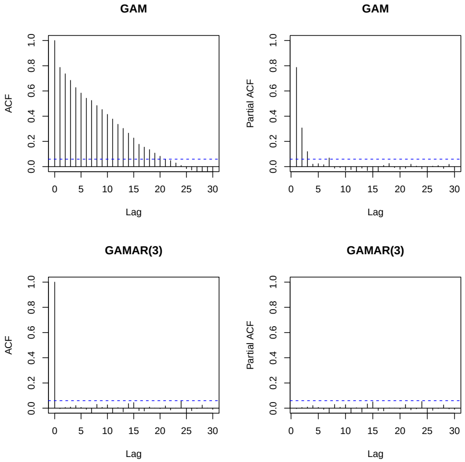

Figure 3: ACF and PACF of GAM and GAMAR (3)

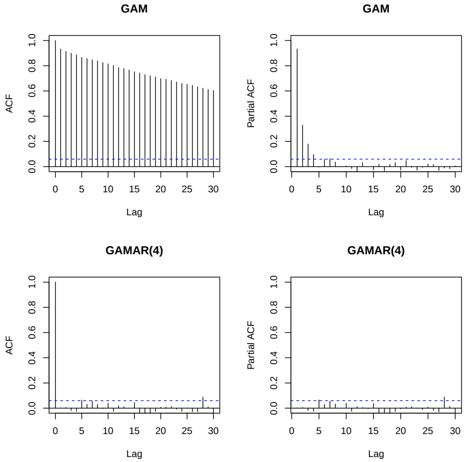

Figure 4: ACF and PACF of GAM and GAMAR (4)

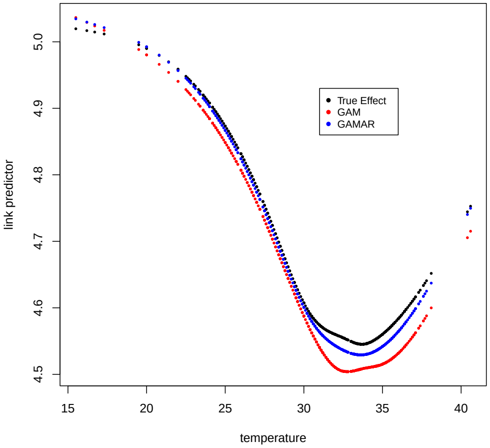

Figure 5: The temperature effects in link scale of GAM and GAMAR (3)

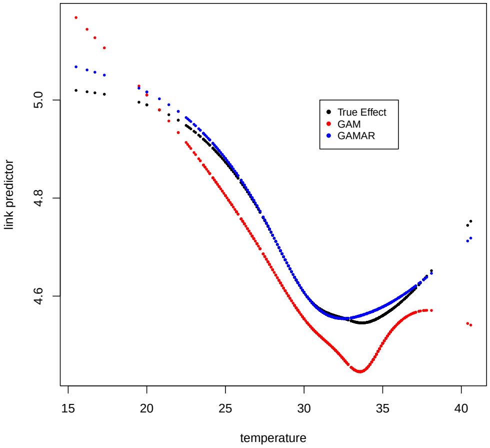

Figure 6: The temperature effects in link scale of GAM and GAMAR (4)

Cases 1-4 ($n = 730$, Lag $= 1, 2, 3$ and 4)

Table 3: Estimates of dispersion parameter

Cases	1	2	3	4
GAM	1.9	2.4	4.5	2.8
GAMAR	1.0	1.1	1.1	1.1

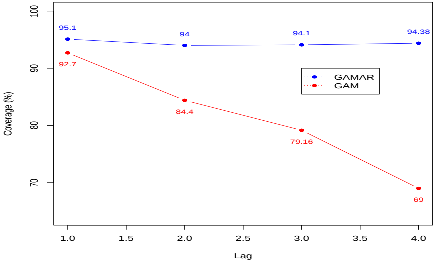

Figure 7: Coverage Plot

Algorithm of Data Generation (Lag 4)

Implementing the following steps yield a single observation from the GAMAR (4):

Step 1: Set the true values $(c_1, c_2, c_3, c_4) = (0.5, 0.25, 0.12, 0.06)$

Step 2: Choose a_t from $a_t \sim \text{Normal}\,(4, 0, 0.2)$

Step 3: Simulate $y_t \sim \text{Poisson}(\mu_t)$ where

- $\ln(\mu_t) = 3.5 + 0.4\cos\Big(0.2\pi\,(x_t + 5)\Big) + \sum_{i=1}^{p} c_i \Big[\ln\big(y^*_{t-i}\big) - \big(3.5 + 0.4\cos\left(0.2\pi\,(x_t + 5)\right)\big)\Big]$
- $y^*_t = \max\,(y_t, \tau)$, $\tau = 0.5$ and x_t = covariate

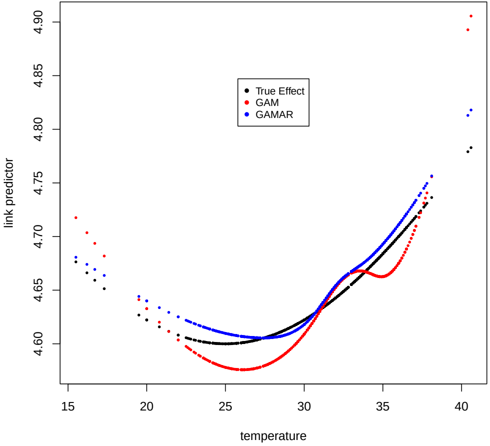

Figure 8: The temperature effects in link scale of GAM and GAMAR (3)

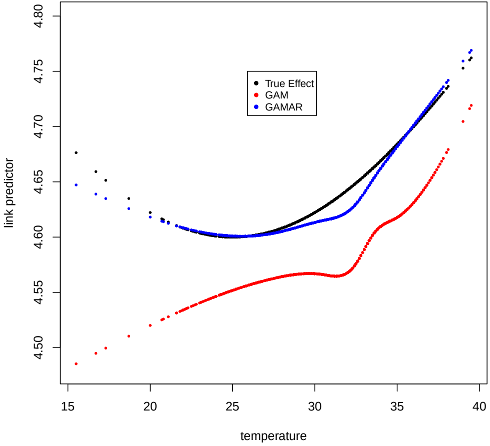

Figure 9: The temperature effects in link scale of GAM and GAMAR (4)

Application to Real Life Data

- **Data:**
  - Directorate General of Health Services (DGHS)
  - Bangladesh Meteorological Department (BMD)
- **Response Variable:** Dengue Infected Cases $(2022 - 2023)$
- **Explanatory Variables:**
  - Average Temperature
  - Rainfall
  - Visibility
  - Wind Speed
  - Humidity

Scatter Plots of Predictors

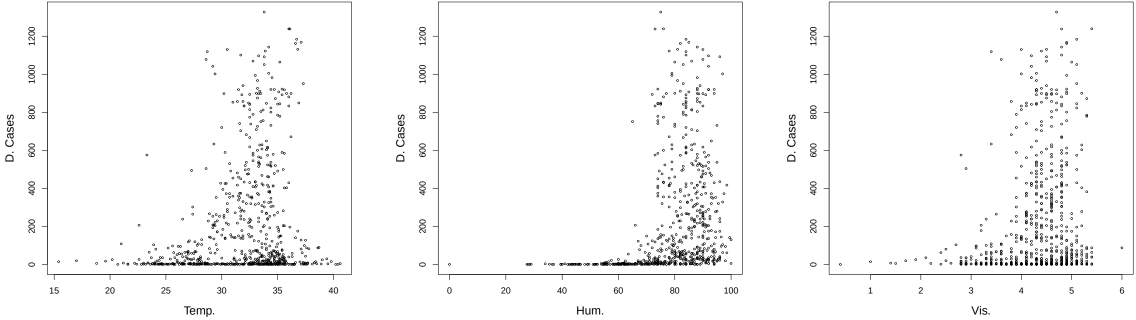

Figure 10: Scatter plots of predictors

Model with Lagged Temperature and Lagged Humidity

For lagged temperature, the model is

$$
\begin{aligned}
\ln\left(\mu_t\right) &= f\left(x_t\right) + \sum_{j=1}^{6} c_j \left(\ln\left(y^*_{t-j}\right) - f\left(x_{t-j}\right)\right), \\
f\left(x_t\right) &= \beta_0 + ns(\text{time}) + ns(\text{temperature}_{t-6}) \\
&\quad + ns(\text{visibility}_t) + ns(\text{wind}_t) + ns(\text{rain}_t) \\
&\quad + ns(\text{humidity}_t) + w_t(\text{week}_t).
\end{aligned}
$$

For lagged humidity, the model is

$$
\begin{aligned}
\ln\left(\mu_t\right) &= f\left(x_t\right) + \sum_{j=1}^{5} c_j \left(\ln\left(y^*_{t-j}\right) - f\left(x_{t-j}\right)\right), \\
f\left(x_t\right) &= \beta_0 + ns(\text{time}) + ns(\text{humidity}_{t-5}) \\
&\quad + ns(\text{visibility}_t) + ns(\text{wind}_t) + ns(\text{rain}_t) \\
&\quad + ns(\text{temperature}_t) + w_t(\text{week}_t).
\end{aligned}
$$

ACF and PACF Plots

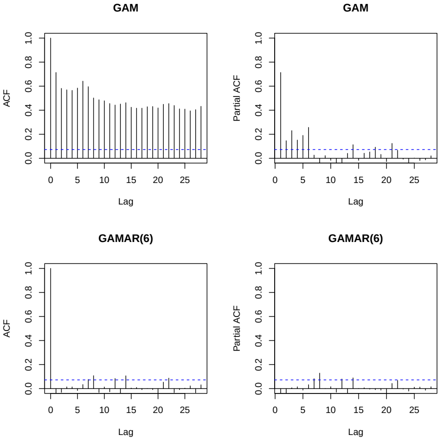

Figure 11: ACF and PACF for temperature

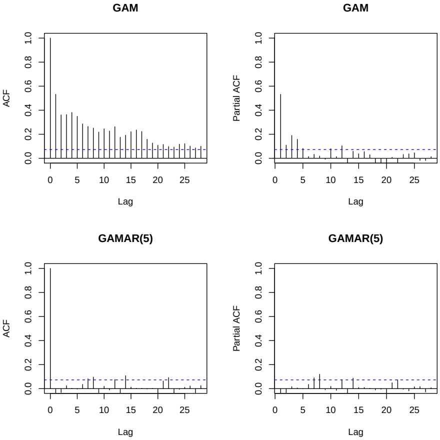

Figure 12: ACF and PACF for humidity

Model Comparison

For lagged temperature

Table 4: Performance Metrics for GAM and GAMAR

Method	AIC	Log-likelihood	Deviance	Dispersion
GAM	130900.6	-65420.28	127137.5	20.20
GAMAR	16553.62	-8238.81	12799.96	1.11

For lagged humidity

Table 5: Performance Metrics for GAM and GAMAR

Method	AIC	Log-likelihood	Deviance	Dispersion
GAM	95179.03	-47542.52	91381.97	20.16
GAMAR	16855.34	-8375.66	13063.18	1.33

Partial Effect of Lagged Temperature

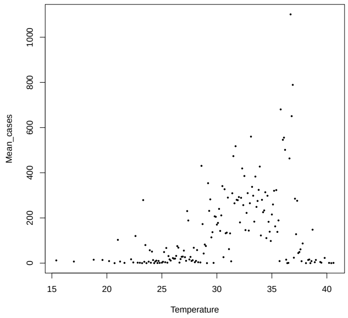

Figure 13: Scatter plot of temp vs mean cases

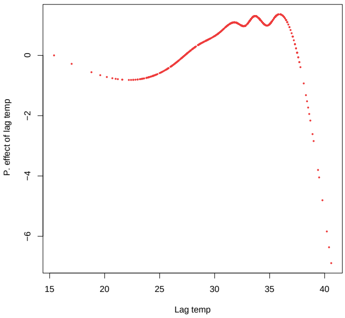

Figure 14: Partial effect plot from GAM

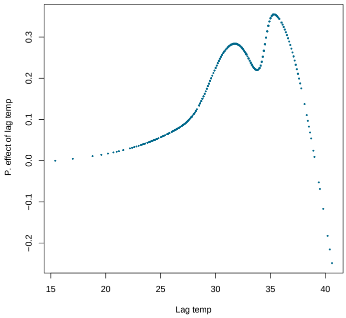

Figure 15: Partial effect plot from GAMAR (6)

Partial Effect of Lagged Humidity

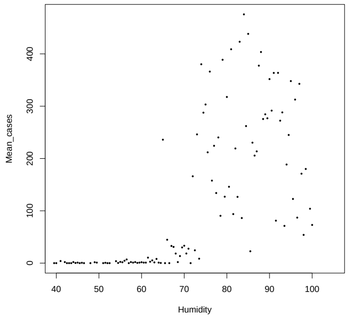

Figure 16: Scatter plot of hum vs mean cases

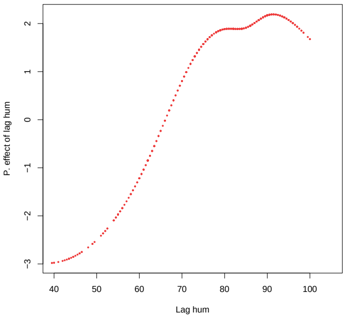

Figure 17: Partial effect plot from GAM

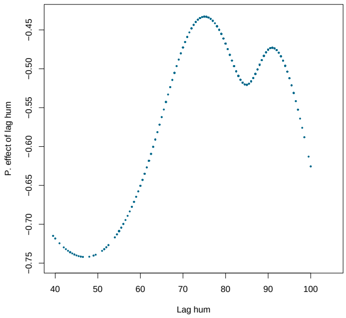

Figure 18: Partial effect plot from GAMAR (5)

Findings and Future Scope

Findings from the study:

- GAMAR outperforms GAM across different lags and functional forms
- GAMAR performs better consistently, regardless of sample sizes and lag length
- GAM's performance deteriorates with increasing lag length
- Temperature and humidity exhibits complex nonlinear relationships with dengue cases

Future Scope:

- Application to other autocorrelated and overdispersed count data
- Explore simulations involving more than one covariates

References

- Yang, L., Qin, G., Zhao, N., Wang, C., Song, G. (2012). Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. BMC medical research methodology, 12, 1-13.
- Hossain, S. (2023). Generalized Linear Regression Model to Determine the Threshold Effects of Climate Variables on Dengue Fever: A Case Study on Bangladesh. Canadian Journal of Infectious Diseases and Medical Microbiology, 2023(1), 2131801.
- Directorate General of Health Services (DGHS), Ministry of Health and Family Welfare, Government of the People's Republic of Bangladesh. Directorate General of Health Services. http://www.dghs.gov.bd.
- Bangladesh Meteorological Department (BMD). Retrieved from http://bmd.gov.bd.