Partial Likelihood Analysis of Autoregressive Models for Disease Incidence Data

Eric V. Slud, Univ. of Maryland, Mathematics Dept.
& Center for Statistical Research & Methodology
U.S. Census Bureau

Univ. of Utah Biostatistics, Dec. 8, 2023

Overview

- Conditional specification of event-count models with time-dependent covariates
- Large-sample theory using Partial Likelihood & martingales
- Causality: Time-dependent covariates influenced by Outcome
- Parametric specifications of various stochastic regression (also called *Markov Regression* or *autoregressive*) models
- Examples of model building involving London Mortality and Dengue fever data
- Value of including past residuals as predictors

Modeling Time-Dependent Outcomes in Terms of Unmodeled Time-dependent Covariates

$Y_t,\ t \geq 1,$ outcome variable such as measured disease incidence

$Z_t,\ t \geq 0,$ vector of predictive variables (risk factors, disease precursors, exposure measurements)

Objective: k-step predictive distribution $f(y_{t+k} \,|\, z_s,\ y_s,\ s \leq t)$

Method: (semi-) parametric model, estimation via partial likelihood

Example: Dengue Fever Incidence

Outcome: **daily new dengue fever cases in specific area**

Predictive variables:

Rainfall, Age of Housing (Village clusters)	Taiwan data
Temperature, (relative) humidity	Taiwan data
Absolute Humidity	Singapore data
'Narrow alleys' measure within village	Taiwan

geographic covariates could serve as proxy for spatial model

Methodological & Theoretical Issues

If $(Z_r,\, r \geq t)$, Y_t conditionally independent given $(Z_s,\, Y_s,\, s < t)$, **for all** t), then $f_{Y_1,\ldots,Y_t \,|\, Z_0,\ldots,Z_{t-1}} = \prod_{j=1}^{t} f_{Y_j \,|\, (Y_s,\, Z_s,\, s<j)}$

This setting extends the no-covariate case to inference for Y on Z regression based on conditional likelihood.

More generally (when Y outcomes can influence later Z's) the product of conditional likelihoods on the right-hand side is not a likelihood or conditional likelihood but is a Partial Likelihood (Cox 1975). **Z's are then also outcome variables**

Example: suppose Mosquito abundance (measured somehow) is a predictive variable and emergency mosquito-control measures are intensified at times when Dengue outbreaks occur, then the PL but not the conditional-likelihood case holds.

Related Model Application

rainfall predicting **runoff** (Slud & Kedem 1994) in hydrology

outcome variable $y_t = I_{[r_t \geq c]}$ indicating that day-t runoff exceeds c, examined for two levels of runoff threshold c

fitted against vectors V_t of lagged rainfall R_t and runoff r_t values, using logistic-regression, likelihood ratio tests, diagnostics and

goodness of fit statistics involving quadratic forms of observed minus predicted counts $\sum_{t:\, V_t \in A_g} (y_t - \widehat{y}_t)$

where $A_g,\ g = 1, \ldots, G$ partitions the space of covariate values

Large-Sample Theory via Martingales

In general case of parametric stochastic regression of Y_t on the past data $\mathcal{F}_{t-1} = (Y_s, Z_s,\ s \le t-1)$ at lags $\le m$:

$\nabla \log PL = \sum_{t=m+1}^{T} \nabla_\theta \log f(Y_t \,|\, Y_s, Z_s,\ s = t-m, \ldots, t-1;\, \theta)$

is a martingale in T

Slud and Kedem (1994, Statist. Sinica) show under regularity conditions that the PL maximizer $\widehat{\theta}$ is Consistent Asymptotically Normal with Wilks-Theorem-type Likelihood Ratio testing using

$$\sqrt{T-m}\,(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\, V)$$

$$\widehat{V} = \Big(- \sum_{t=m+1}^{T} \nabla_\theta \nabla_\theta' \log f(Y_t \,|\, \mathcal{F}_{t-1},\, \theta) \Big|_{\theta=\widehat{\theta}} \Big)^{-1}$$

Continuous-time extended PL theory in Slud (1992, Scand. J. Stat.)

Examples of Autoregressive Conditional Models

In spirit of Zeger & Qaqish(1988), Slud & Kedem (1994) ...

$Y_t \,|\, \mathcal{F}_{t-1}$ may depend on $\mathbf{Y}_{t-1}^m = (Y_{t-1}, \ldots, Y_{t-m})$, $\mathbf{Z}_t$ through

linear ARMAX $Y_t = \beta' \mathbf{Y}_{t-1}^k + \gamma' \mathbf{Z}_{t-1}^k + e_t\,, \quad E(e_t | \mathcal{F}_{t-1}) = 0$

Poisson regression $Y_t \sim \text{Poisson}(\lambda_t),\ \log \lambda_t = \beta' \mathbf{Y}_{t-1}^k + \gamma' \mathbf{Z}_{t-1}^k$

logistic regression $P(Y_t = 1 \,|\, \mathcal{F}_{t-1}) = \texttt{plogis}(\beta' \mathbf{Y}^{*k}_{t-1} + \gamma' \mathbf{Z}_{t-1}^k)$

with recoded $\mathbf{Y}^*$ involving past residuals

linear predictive scores $\beta' \mathbf{Y}^{*k}_{t-1} + \gamma' \mathbf{Z}_{t-1}^k$ may also have form of distributed lags (e.g., coeff's parameterized in decaying pattern)

$$\sum_{j=1}^{k} \left(\beta_j(\alpha)\, Y^*_{t-j} + \gamma_j(\rho)\, Z_{t-j} \right)$$

Modeling the London Mortality Data

City mortality-count models inspired by time-series methods:

LA mortality data (smog) (Shumway et al. 1988),
by state-space time series methods

London mortality (temp, humidity, smoke & pollutants)
(Schwartz & Marcus 1990), with additive AR year-effects

London Mortality data, 112 winter days 1958–59 to 1971–72

outcome: daily all-cause mortality counts

predictors: 'British smoke', sulfur dioxide, temp, humidity

Model Building Steps for London Mortality Data

(1) hypothetical individual-level Cox model

if environmental covariates V_t and subject-specific covariates Z_{it} for subject i born on day b_i were available, Cox Proportional Hazards model might say

$$\text{failure rate on day } t \;=\; \lambda_0(t - b_i)\exp(\beta' Z_{ti} + \gamma' V_t)$$

Rate with subject covariates **not** observable $= c \cdot e^{\gamma' V_t}$

(2) aggregated Poisson model for n_t = number dying on day t

individual daily failure rates are tiny; aggregate rare successes from many independent coin-tosses to get approximately Poisson variates with rate $E(n_t) = \alpha \exp(\gamma' V_t)$

Model Building Steps, continued

(3) linear regression on time (within-year, or overall)

Demographers model linear secular changes in mortality rates by multiplicative factor $a+bt$ or, within year $y = y(t)$, by $a_y+b_y\,t$ or exponential, leading to Poisson regression model

$$n_t \;=\; N(t) - N(t-1) \;\approx\; \text{Poisson}(\exp\{a_y + b_y t + \gamma' V_t\})$$

where $N(t)$ denotes cumulative mortality through day t.

(4) other interactions between subject and environment

with interaction terms $\delta\, V_t\, Z_{ti}$ in Cox-model exponent, averaging could give multiplicative factor $g(V_t, \delta)$, e.g., sicker subjects susceptible to cumulative environmental risk factors.

Discussion of Residuals plots and Diagnostics

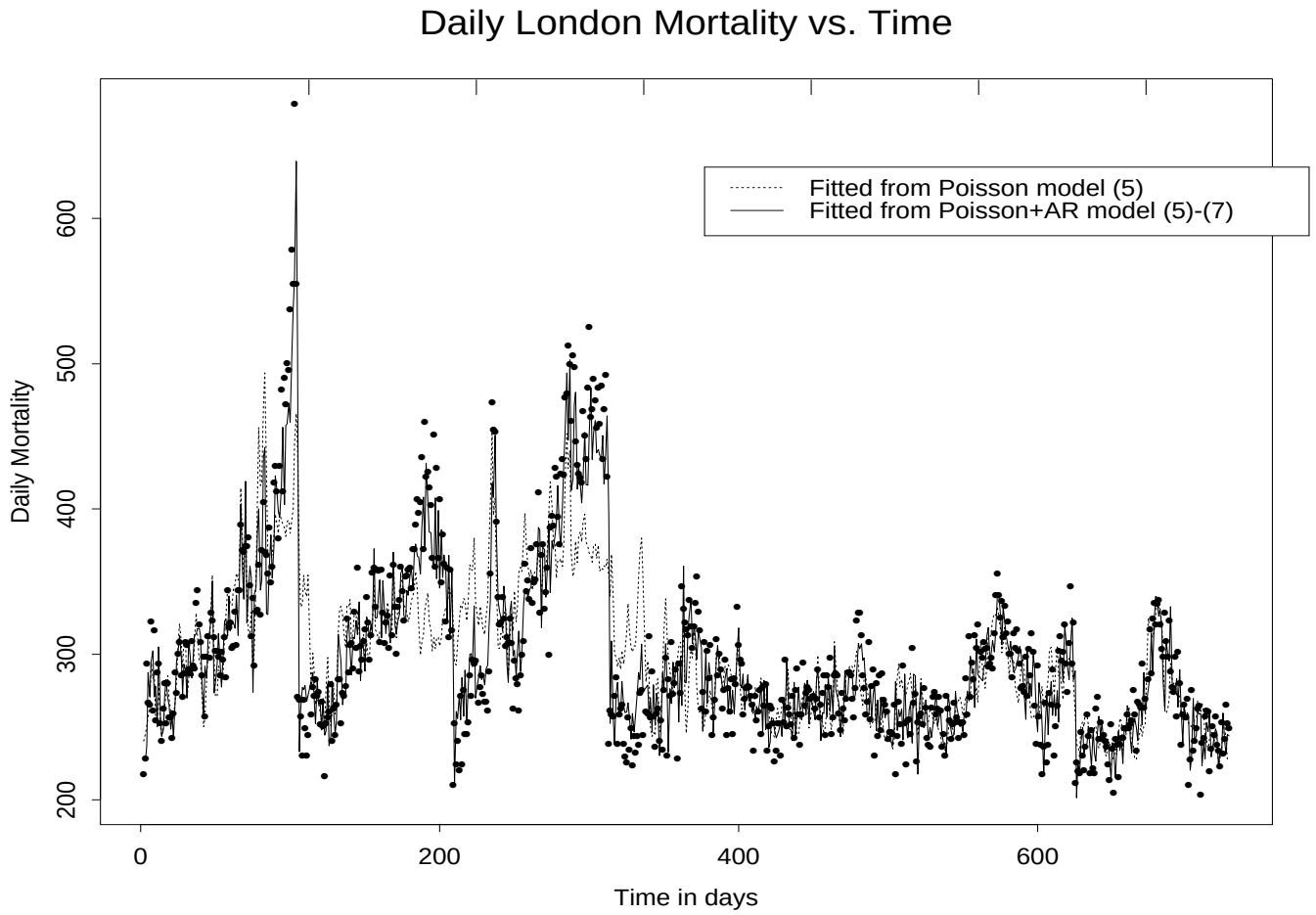
Recoded and lagged covariates

$$V_t = (\sqrt{\text{Smok}_t}, \sqrt{\text{Smok}_{t-1}}, \sqrt{SO2_t}, \sqrt{SO2_{t-1}}, \text{Tmp}_t, \text{Tmp}_{t-1})$$

Residuals $\Delta_t = n_t - \widehat{n}_t$ from Poisson regression model analyzed through residual-autroregression model $\Delta_t = c_0 + c_1 \Delta_{t-1} + \epsilon_t$

Corresponding mortality-versus-predictor plots in the following figure.

Over-plotted mortality and fits with and without residual autoregression.



Plotted daily London Mortality versus enivironmental and weather predictors, with and without residuals autoregression

Dengue Fever Incidence Models

Model publicly available numbers of daily new Taiwan Dengue cases Y_t in 2014 and 2015 in terms of (lagged) Rainfall R_s and cases $Y_s,\ s < t$.

Reasoning as in London Mortality analysis, specify Y_t given $\mathcal{F}_{t-1} = \{Y_s, R_s,\ s < t\}$ as Poisson(λ_t).

Mechanism for Dengue following rainfall is mosquito breeding: the effect of rainfall is like lagged impulse-response, so think of

$$\lambda_t = \sum_{s:\, s<t} R_s\, \varphi_{t-s}$$

for some function $\varphi_j = \varphi(j,\, \theta)$ of lags j and parameters θ.

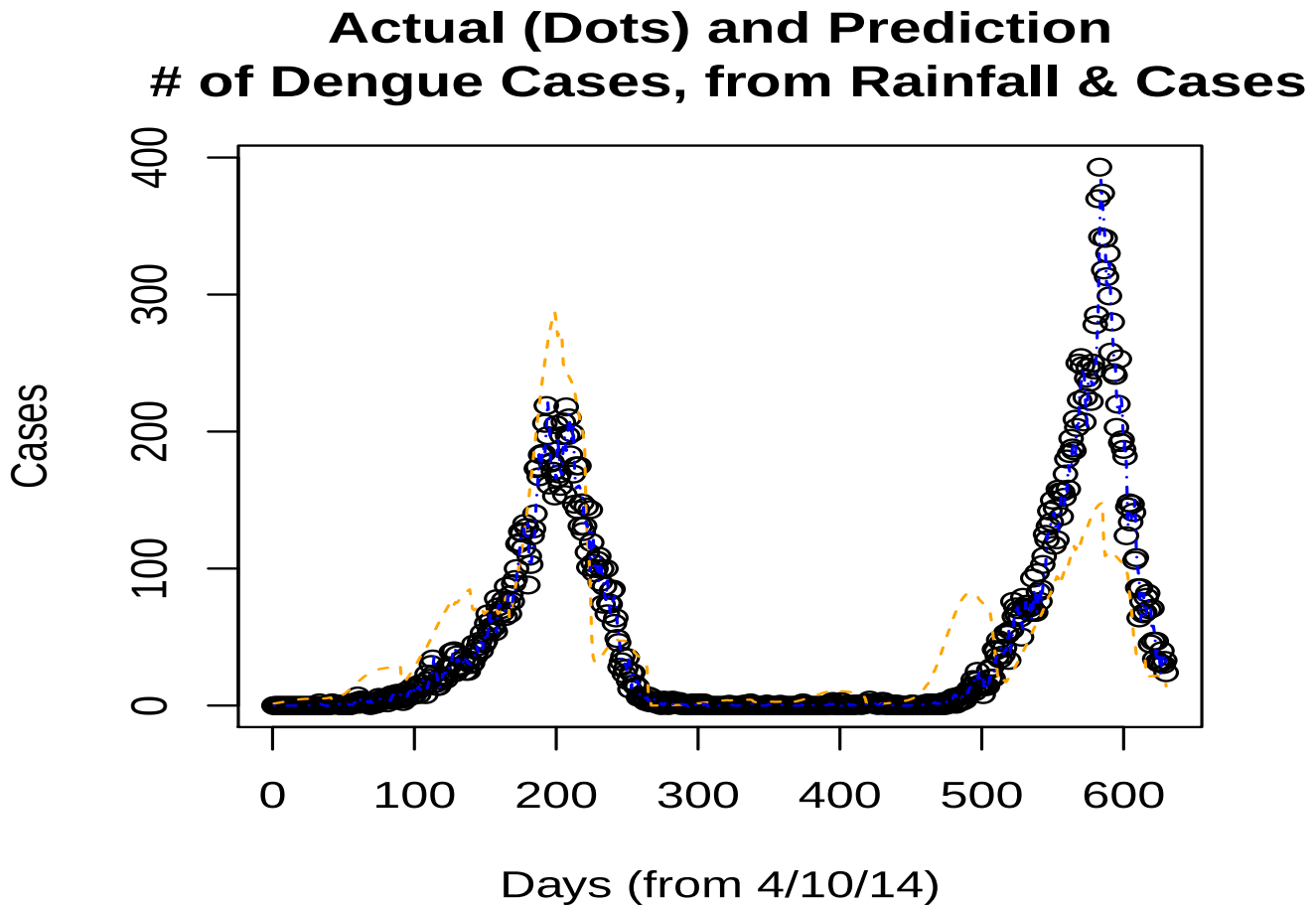
Further Dengue Model Formulation

Idea of distributed lag model well known in econometrics, also used in Singapore paper Xu et al. (2014 PLOS Neglected tropical Diseases), with Absolute Humidity as covariate. Other time-dependent covariates tried in Taiwan Dengue modeling include: (Absolute) humidity, Age of housing, Narrow Alleys, or Village and geography identifiers.

In model fits below (with Rainfall alone), φ_j is proportional to Gamma distribution masses on integer intervals $(j-1, j]$

$$\varphi_j \;=\; \alpha \cdot \Big(\texttt{pgamma}(j, \mu, \lambda) - \texttt{pgamma}(j-1, \mu, \lambda) \Big)$$

with 3 free parameters $\theta = (\alpha, \mu, \lambda)$.



Daily 2014 Taiwan Dengue Cases vs. Rainfall, two predictive models (orange without residuals autoregression, blue with, and $\text{cor}(n_t, \widehat{n}_t) = 0.988$.)

Dengue Modeling Steps

Taiwanese modeling team: restricted attention to Kaohsiung city data; predicted 5 and 15 days ahead, and abandoned attempt at parsimony using distributed lag parameterization. Models are:

$$Y_t = \alpha + \sum_{j=1}^{D} \beta_j \, Y_{t-j} + \sum_{j=1}^{D} \gamma_j \, H_{t-j} + \sum_{j=1}^{D} \delta_j \, BR_{t-j} \qquad \textbf{(M1)}$$

$$Y_t = \textbf{same} + \sum_{j=1}^{D} \varphi_j \, \widehat{Y}_{t-j}^{(M1)} \cdot \mathrm{sgn}(\widehat{Y}_{t-j}^{(M1)} - 100) \qquad \textbf{(M2)}$$

where $D \approx 50$, and where 100 is a large number of daily cases for Kaohsiung.

Summary

The talk described a generalized-linear model building strategy that follows usual (Poisson) regression steps based on time-dependent and lagged covariates — transforming covariates, diagnostic plotting, likelihood-ratio-test and AIC tools —

considers distributed-lag parameterization of effects of environmental variables

and tries lagged residuals as additional covariates that have worked well in previous examples and preliminary analysis.

References

Cheng, Y-C,...,Slud, E.,...,Tsou, Hsiao-Hui (2020, PLOS *Neglected Trop. Diseas.*, 'Real-time Dengue Forecasting...'

Paul Albert papers with various co-authors

Schwartz & Marcus (1990), Slud (1997) London Mortality Data

Shumway, Azari & Pawitan (1988) L.A. mortality data

Slud, E. and Kedem, B. (1994) Statistica Sinica

Zeger, S. and Qaqish, B. (1988, Biometrics) Markov regression

Rainfall-Runoff, continued

Since runoff (almost) does not affect future rainfall: **conditional likelihoods** given all Rainfall data, would suffice for rainfall-runoff models if the data are $V_t = (r_t,\, R_t),\ \ t \geq 0$,

but if daily measurements are intermittent

$$V_t \equiv (r_t,\, R_t),\ \ t = 1, 2, 4, 7, 8, 12, \ldots$$

then runoff value R_7 **does** give information about missed rainfall measurements r_5, r_6, so : $(r_t,\ t \geq 7)$ is not conditionally independent of R_7 given V_1, V_2, V_4 (even though it would be if we conditioned also on V_3, V_5, V_6).

Similarly for Latent Health-Status modeling

(related to current research with Dan Scharfstein)

Suppose Z_t denotes underlying health-status, Y_t medical measurements, T_t treatment intervention indicators, then

$$Z_t \Rightarrow Y_{t+1} \Rightarrow T_{t+2} \Rightarrow Z_{t+3}$$

So if we model Z_t, Y_t, we must use **partial** likelihood $f_{Y_t \,|\, (Z_s,\, s<t)}$ not **conditional** likelihood.

Thank you !

Eric.V.Slud@census.gov